PRO GRADU -TUTKIELMA

Saara Oinonen

Genominlaajuisen SNP-aineiston tutkimusmenetelmien vertailua syöpää aiheuttavien perimän muutosten löytämiseksi

TAMPEREEN YLIOPISTO
Informaatiotieteiden yksikkö
Tilastotiede

Kesäkuu 2011

Tampereen yliopisto
Informaatiotieteiden yksikkö
OINONEN, SAARA: Genominlaajuisen SNP-aineiston tutkimusmenetelmien vertailua syöpää aiheuttavien perimän muutosten löytämiseksi
Pro gradu -tutkielma, 34 s., 2 liites.
Tilastotiede
Kesäkuu 2011

Tiivistelmä

Tämän tutkielman tarkoitus on kokeilla ja vertailla erilaisia laajan geneettisen aineiston tutkimiseen tarkoitettuja tilastollisia menetelmiä. Huomiota kiinnitetään eri analyysimenetelmien antamien tulosten samankaltaisuuteen ja siihen, kuinka hyvin ne sopivat aineiston analysointiin. Geneettiset aineistot ovat tilastotieteen kannalta haastavia tutkittavia, sillä muuttujat ovat usein toisistaan riippuvaisia ja niitä saattaa olla aineistossa satoja tuhansia. Tässä tutkielmassa onkin tarkoitus arvioida muutamaa perinteisempää menetelmää sekä uudempia, genetiikan tutkimuksille varta vasten kehitettyjä menetelmiä. Tarjolla olevista lukuisista menetelmistä tässä tutkielmassa käytetään χ^2-riippumattomuustestiä, logistista regressioanalyysia, kytkentäepätasapainoanalyysia ja haplotyyppiblokkien määrittämistä sekä *Random forest* -algoritmia.

Avainsanat: χ^2-riippumattomuustesti, logistinen regressioanalyysi, kytkentäepätasapaino, haplotyyppiblokit, Random forest.

Sisältö

1 Johdanto

DNA:n rakenteen määrittäminen vuonna 1953 oli tieteellinen läpimurto (Watson & Crick 1953), jonka seurauksena monimuotoisten sairauksien syntymisen syitä voitiin alkaa etsimään ihmisen perimästä. Kuluneen 60 vuoden aikana menetelmät geneettisen koodin selvittämiseksi ovat kehittyneet huimasti, ja ihmisen genomi eli perimä saatiin selville vuonna 2003 (Collins, Morgan & Patrinos 2003). Nykyään laajoja, koko genomin kattavia geneettisiä aineistoja on paljon, mutta sopivat tutkimusmenetelmät, joilla aineistot saataisiin hyödynnettyä mahdollisimman hyvin, ovat vielä kehitysasteella. Monimuotoisten sairauksien, kuten syöpien, tutkimiseen räätälöityjä testejä ja algoritmeja on paljon. Usein eri menetelmien antamat tulokset ovat kuitenkin ristiriidassa keskenään sekä aiemmin saatuihin tuloksiin nähden ja siksi menetelmien vertaileminen ja edelleen kehittäminen on tärkeää. (Cantor, Lange & Sinsheimer 2010.)

SNP-markkerit eli yhden emäksen polymorfismit ovat suosittuja tutkimuskohteita mm. syöpäsairauksien tutkimuksissa, sillä ne kattavat koko genomin ja aineistoja on helposti saatavilla tietokannoista. SNP-markkereiden avulla voidaan genomista paikantaa alueita, joilla on merkitystä perinnöllisten tautien syntymiseen. SNP-markkerit ovat haastava tutkimuskohde tilastotieteen kannalta, sillä monien perinteisten menetelmien oletukset, kuten muuttujien riippumattomuus, eivät SNP-aineistojen kohdalla päde.

Tässä tutkielmassa aineistona on noin 91 000 SNP-markkerin rintasyöpäaineisto, jossa on 84 syöpätapausta ja 393 verrokkia. Rintasyöpätapaukset ovat Tampereen yliopiston Biolääketieteellisen teknologian instituutin keräämiä yhteistyössä Tampereen yliopistollisen sairaalan kanssa ja verrokkiryhmä on saatu pohjoismaisesta yhteistyöprojektista *Nordic Centre of Excellence in Disease Genetics* (http://www.ncoedg.org/). Tavoitteena on löytää sellainen analyysimenetelmä, jonka avulla voidaan kartoittaa ne alueet genomista, joilla voi olla vaikutusta syövän syntymiseen. Analyysissä on tarkoitus ottaa huomioon SNP-markkereiden vaikutus sekä yksittäin että ryhminä.

Aineiston laajuus asettaa monia haasteita niin käytettäville menetelmille kuin tietokoneiden laskentatehokkuudelle. SNP-markkerit testattiin ensin yksittäin käyttäen χ^2-testiä sekä logistista regressioanalyysiä. Samalla tarkasteltiin testien antamien tulosten yhdenmukaisuutta ja monimuuttujaisen aineiston yhteydessä yleistä hylkäys-virhettä (*false discovery rate*). Ryhmittelyyn käytettiin kytkentäepätasapainomääritelmää, jonka avulla SNP-markkerit ryhmiteltiin haplotyyppiblokkeihin. Haplotyyppiblokkien vaikutusta henkilön sairastuvuuteen testattiin myös logistisella regressioanalyysillä. Haplotyyppiblokeista piirrettiin lisäksi kuvat, joiden avulla nähdään blokkien sijoittuminen kromosomeissa.

SNP-markkerit ryhmiteltiin vielä *Random forest* -menetelmällä, joka on Breimanin (2001) kehittämä nopea ja tehokas koneoppimiseen perustuva luokittelualgoritmi. *Random forest* pyrkii löytämään aineistosta sellaiset muuttujat, jotka jakavat havainnot parhaiten syöpää sairastavien ja terveiden kesken. Menetelmä sopii hyvin tässä tutkielmassa käytetyn aineiston tapaisille laajoille aineistoille, sillä SNP-markkereiden väliset riippuvuudet eivät ole algoritmille ongelma. Lopuksi kaikkien menetelmien tuloksia vertaillaan keskenään ja arvioidaan, onko jokin menetelmistä muita parempi.

2 Aineiston kuvaus

Ihmisen perimä on tallennettu solun tumaan kromosomeihin DNA-ketjuksi, joka koostuu sokeri-fosfaattirangasta ja neljän emäksen, adeniinin, tymiinin, guaniinin ja sytosiinin, muodostamasta geneettisestä koodista. Näiden emästen keskinäinen järjestys määrää geenien ilmentymän. Perimän muutoksia sanotaan mutaatioiksi ja oletetaan, että näillä mutaatioilla on yhteys monimuotoisten sairauksien syntymiseen.

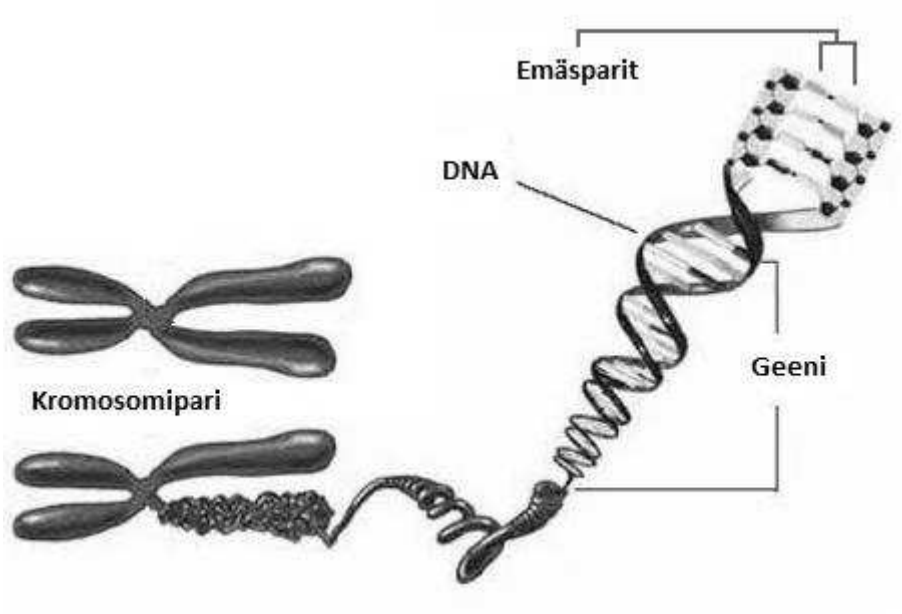
Kirjainyhdistelmä SNP tulee englanninkielen sanoista *single nucleotide polymorphism*, ja se voidaan suomentaa yhden emäksen monimuotoisuudeksi. SNP on siis yhden emäksen pistemutaatio, joiden avulla voidaan tutkia perimän eroavaisuuksia ihmisten ja populaatioiden välillä. Ne ovat suosittuja tutkimuskohteita mm. syöpäsairauksien syntymekanismin määrittämisessä, sillä ne kattavat koko genomin ja nykytekniikalla ne ovat helposti mitattavissa.

Tutkielman aineistona on genominlaajuinen rintasyöpäaineisto, jossa on noin 91 000 SNP-muuttujaa. Aineistossa on 84 rintasyöpätapausta ja 393 verrokkia, joista kaikki ovat naisia. Rintasyöpätapaukset ovat Tampereen yliopiston Biolääketieteellisen teknologian instituutin keräämiä yhteistyössä Tampereen yliopistollisen sairaalan kanssa ja verrokkiryhmä on saatu pohjoismaisesta yhteistyöprojektista *Nordic Centre of Excellence in Disease Genetics* (http://www.ncoedg.org/).

2.1 Ihmisen genomi ja SNP-markkerit

Ihmisen geneettinen koodi on kirjattu solujen tuman kromosomeihin, joita ihmisellä on 23 paria. Kromosomiparista toinen on periytynyt isältä ja toinen äidiltä. Kromosomeihin pakattu DNA-ketju koostuu fosfaatti-sokerirangasta, johon on kiinnittynyt emäksiä. Nämä neljä eri emästä, adeniini (A), tymiini (T), guaniini (G) ja sytosiini (C), muodostavat geneettisen koodin. DNA:n kaksois-kierteistä rakennetta kutsutaan heliksiksi. Kaksijuosteisessa DNA-heliksissä kahden ketjun emäkset ovat kiinnittyneet toisiinsa niin, että adeniini saa parikseen aina tymiinin ja guaniini sytosiinin. Molemmissa ketjuissa on sama geneettinen koodi, mutta ne ovat toistensa komplementteja. Koska kromosomeja on parillinen määrä, on ihmisellä yhteensä neljä kopiota samasta geenistä (lukuun ottamatta suku-puolikromosomeja X ja Y). (Engle, Simpson & Landers 2006.)

Kuvassa 1 on havainnollistettu perimän rakennetta. Kromosomiparit ovat toistensa kaltaisia, sillä ne sisältävät samat geenit. Geenien koodi saattaa kuitenkin vaihdella kromosomien kesken, sillä parista toinen on peritty isältä ja toinen äidiltä. DNA on pakattu kromosomiin kaksoiskierteenä eli heliksinä, jonka juosteet ovat keskenään identtisiä, mutta komplementteja. Toista juostetta luetaan siis oikealta vasemmalle ja vastakkaista vasemmalta oikealle. Geenit ovat osana DNA-juostetta, joten yhdessä kaksoiskierteessä on geenistä kaksi identtistä kopiota. Koska kromosomiparin kromosomit sisältävät samat geenit, on yhdestä geenistä yhteensä neljä kopiota.



Kuva 1 DNA ja geenit kromosomissa.

Ihmisten genomit ovat jopa noin 99,9 % samanlaisia. Geneettisillä markkereilla tarkoitetaan genomien välisiä eroja, joiden avulla voidaan tutkia genomin ilmentymien eroavaisuuksia yksilöiden tai populaatioiden välillä. Tällaisia markkereita ovat myös SNP-markkerit. SNP, joka on lyhenne englanninkielen sanoista *single nucleotide polymorphism*, voidaan suomentaa yhden emäksen polymorfismiksi eli monimuotoisuudeksi. SNP-markkerit ovat siis perinnöllisiä yhden emäksen mutaatioita eli pistemutaatioita, jotka mitataan henkilön molemmista kromosomiparin kromosomeista. Saman SNP-markkerin eri muotoja sanotaan alleeleiksi. (Engle et al. 2006.) Kuvassa 2 on aukaistu kromosomiparin DNA-ketjuja samasta kohtaa, ja SNP on merkitty nuolella. Kuvan SNP-markkerin alleeliksi tulisi siis joko AC tai TG, riippuen kummalta juosteelta emäkset luetaan. Useamman SNP-markkerin muodostamaa emästen ryhmää (kuten AAGACC) sanotaan haplotyypiksi. SNP:n sijaintia DNA-ketjussa sanotaan lokukseksi, joka määritetään emäs-parin (bp, *base pair*) järjestysnumeron perusteella. Esimerkiksi kuvan 2 SNP:n lokus on 28.

A	C	T	T	C	G	A	G	C	C	G	A	A	T	A	T	G	C	C	G	C	A	T	A	A	T	T	**A**	C	G	
T	G	A	A	G	C	T	C	G	G	C	T	T	A	T	A	C	G	G	C	G	T	A	T	T	A	A	**T**	G	C	
1	bp													15																30
A	C	T	T	C	G	A	G	C	C	G	A	A	T	A	T	G	C	C	G	C	A	T	A	A	T	T	**C**	C	G	
T	G	A	A	G	C	T	C	G	G	C	T	T	A	T	A	C	G	G	C	G	T	A	T	T	A	A	**G**	G	C	

Kuva 2. SNP:n määrittäminen DNA-ketjusta.

Nykyään käytössä ovat genominlaajuiset SNP-markkerisirut, joilla voidaan samanaikaisesti analysoida miljoonia SNP-markkereita tuhansista yksilöistä. Tämä mahdollistaa geneettisen epidemiologisen tutkimuksen sairauksissa, joissa sairauden aiheuttajana ei ole vain yksi ge-

neettinen muutos. Nykytietämyksen mukaan mm. syöpä on sairaus, joka johtuu useista perimän muutoksista yhdistettynä ympäristötekijöihin. (Engle et al. 2006.)

2.2 Geenikartoitus ja monimuotoisten sairauksien genetiikka

Ihmisen geneettinen koodi saatiin kartoitettua kokonaisuudessaan *Human Genome Projectissa* vuonna 2003 ja se käsittää koko $3*10^9$ pituisen DNA-ketjun emäkset. Koko ketjussa koodaavia geenejä arvellaan olevan n. 25 000 ja geenien etsintä on yhä käynnissä oleva projekti. Yksilön genomia sanotaan genotyypiksi ja geenien ilmentymää, kuten syöpää, fenotyypiksi. SNP-markkerit ovat merkittävä apu uusien geenien löytämisessä sekä perinnöllisten sairauksien kartoittamisessa, sillä niiden avulla koko genomi saadaan tiivistettyä satoihin tuhansiin alleeleihin. Siten on helpompaa paikantaa ne kohdat DNA-ketjussa, joilla on merkitystä esimerkiksi syövän syntyyn. (Botstein & Risch 2003.)

Geenikartoituksen avulla pyritään löytämään tilastollinen yhteys geneettisen markkerin ja ilmiasun välille ja selvittämään, kuinka suuri osuus ilmiasusta voidaan selittää geneettisillä tekijöillä. Geenikartoitukseen on olemassa pääasiassa kaksi lähestymistapaa: kytkentäanalyysi ja assosiaatioanalyysi. Kytkentäanalyysissä tarkastellaan sairauden periytymistä perheaineistossa. Kytkentäepätasapainoon perustuvassa assosiaatioanalyysissa aineisto kerätään väestöstä, jossa sairautta esiintyy. (Ollikainen & Uimari 2004.)

Perinnölliset sairaudet voivat olla joko monogeenisiä, eli yhden geenimuunnoksen aiheuttamia, tai polygeenisiä eli monitekijäisiä. Kun kyseessä on syövän kaltainen monitekijäinen sairaus, mikään geeni yksinään ei riitä selittämään yksilön sairastumista vaan perimän muutoksia tarvitaan useita. Sanotaan, että monitekijäisissä sairauksissa geneettisen markkerin vaikuttavuus (*effect size)* on pieni eli populaatiossa tietyn genotyypin kantajista vain osa ilmentää kytkeytynyttä ilmiasua (epätäydellinen *penetranssi).* Monogeenisen taudin ollessa kyseessä genotyypin vaikuttavuus on suuri ja penetranssi lähes täydellinen eli jopa kaikki tietyn mutaation kantajat ilmentävät siihen assosioitunutta ilmiasua. (Engle et al. 2006.)

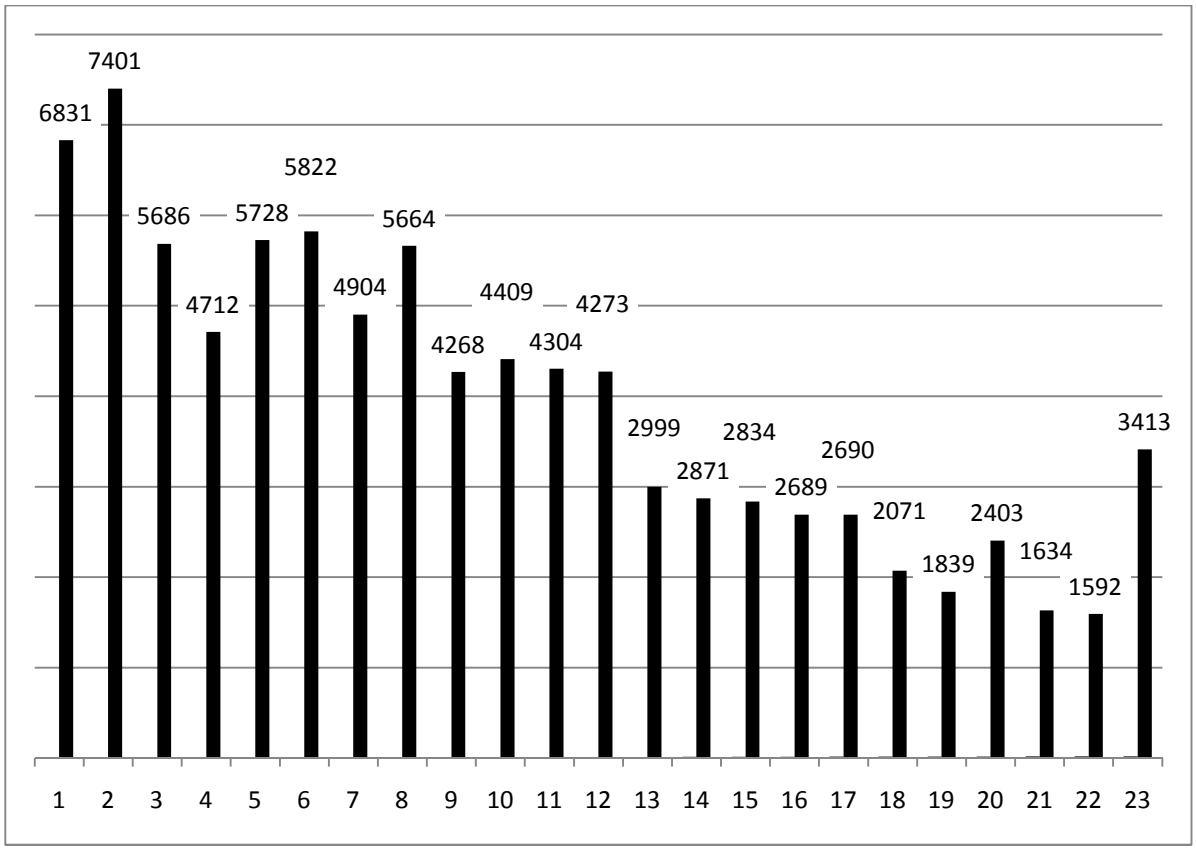
2.3 Aineisto

Tämän tutkielman aineisto koostuu kahdesta eri aineistosta siten, että syöpätapaukset ja kontrollitapaukset ovat eri lähteistä. Rintasyöpätapaukset ovat yhteistyössä Tampereen yliopiston Biolääketieteellisen teknologian instituutin ja Tampereen yliopistollisen sairaalan keräämiä ja näytteet on otettu suomalaisilta rintasyöpäpotilailta. Verrokkiryhmä on saatu käyttöön pohjoismaisesta yhteistyöprojektista *Nordic Centre of Excellence in Disease Genetics* (http://www.ncoedg.org/). Myös kontrollitapaukset ovat suomalaisia. Lopullisessa aineistossa rintasyöpätapauksia on 84 yksilöä ja verrokkeja 393 yksilöä. Kaikki aineiston yksilöt ovat naisia.

Alun perin SNP-markkereita oli tarkoitus analysoida noin 300 000, perustuen syöpätapauksien genotyypin määrittämisessä käytettyyn SNP-siruun (HumanCytoSNP12, Illumina, Inc, San Diego, CA, USA). Mutta sairastuneiden ja verrokkiryhmän SNP-markkereiden yhtäläisyys oli toivottua huonompi, johtuen verrokkien genotyypin määrityksessä käytetyn sirun (Affymetrix 550K, Santa Clara, CA, USA) ja HumanCytoSNP12 sirun välisistä markkereiden eroista. Vaikka verrokkiaineistossa oli markkereita moninkertainen määrä (noin 550 000) syö-

pätapausten markkereihin nähden, lopulliseen aineistoon markkereista päätyi alle kolmannes eli noin 91 000.

Aineistojen yhteensovittamisessa ilmeni myös useita ongelmia, jotka oli ratkaistava ennen varsinaisia analyysejä. Kun SNP-aineistoa kootaan kahdesta eri lähteestä, on otettava huomioon monia sellaisia asioita, jotka voidaan jättää huomiotta, jos koko aineiston markkerit olisivat määritetty samalla menetelmällä. Koska DNA-ketju on kaksijuosteista, on oltava tarkka siitä, että molemmat SNP-aineistot on mitattu samansuuntaiselta juosteelta. Muutoin SNP:t ovat jo lähtökohtaisesti vastakkaisia ja tulokset vääriä.



Kuva 3. SNP-markkereiden jakautuminen kromosomeihin. Vaaka-akselilla on kromosomin numero ja pylväiden päässä on SNP-markkereiden tarkka lukumäärä.

SNP-markkereiden lukumäärä lopullisessa aineistossa on 91 037 ja niitä on koko genomin laajuudelta. Jokaisella SNP-markkerilla on 1–3 alleelia. Kuvasta 3 voidaan tarkastella, kuinka SNP-markkerit ovat jakautuneet eri kromosomeille. Eniten markkereita on kromosomissa 2 ja vähiten kromosomissa 22. Keskimäärin markkereita on noin 3 960 yhdessä kromosomissa. Aineiston ulkopuolelle jääneitä markkereita varten oli tarkoitus hankkia lisää kontrollidataa, mutta niitä ei otettu mukaan tämän tutkielman analyysiin, sillä eri lähteiden käyttö voisi aiheuttaa harhaa tuloksiin.

3 Analyysimenetelmät

Koko genomin kattavia SNP-aineistoja on ollut käytössä jo vuosia ja uusia menetelmiä aineistojen analysointiin kehitetään jatkuvasti. SNP-markkerit ovat tilastotieteen kannalta haastavia tutkittavia, sillä useiden perinteisten menetelmien oletukset, kuten tarkasteltavien muuttujien riippumattomuus, eivät SNP-markkereiden kohdalla päde. Lisäksi DNA:n tutkimusmenetelmien kehittyminen on johtanut siihen, että SNP-aineistot ovat nykyään valtavia. Tällaisia aineistoja varten täytyy siis kehittää täysin uusia analysointimenetelmiä tai ainakin käytössä olevia on muunneltava aineiston tarpeita varten.

SNP-aineistojen analysoinnissa käytetään kuitenkin yhä klassisia tilastollisia analyysimenetelmiä, kuten χ^2-testiä ja logistista regressioanalyysiä. Vaikka nämä menetelmät eivät huomioi SNP-markkereiden keskinäisiä riippuvuussuhteita, voidaan niiden avulla saada yleisnäkemys aineiston lisätutkimuksia tarvitsevista alueista. Suuretkin aineistot voidaan analysoida χ^2-testillä nopeasti, mutta tuloksia on syytä tarkastella kriittisesti ns. väärien positiivisten tulosten takia. Tässä tutkielmassa hylkäysvirhettä arvioitiin käyttäen Benjaminin ja Hochbergin (1995) kehittämää *FDR*-kontrollia.

SNP-markkereiden keskinäisen riippuvuuden tarkasteluun on monenlaisia lähestymistapoja. Datan louhinta ja muuttujien klusterointi ovat laajalti käytettyjä menetelmiä, joilla samantyyppisiä markkereita jaetaan ryhmiksi. (Cantor et al. 2010.) Samantapainen ryhmittelevä menetelmä on SNP-markkereiden kytkentäepätasapainoon perustuva haplotyyppiblokkien muodostaminen. Siinä vierekkäisille SNP-markkereille lasketaan kytkentäepätasapainomitta, joka kuvaa markkereiden välisen kytkennän voimakkuutta. Voimakkaassa kytkennässä olevat markkerit jaetaan blokkeihin eli ryhmiin, jolloin muodostuu monen markkerin alleelien yhdistelmiä eli haplotyyppejä. Näiden haplotyyppien vaikutusta sairauden syntymiseen voidaan testata perinteisin menetelmin, kuten logistisella regressioanalyysilla. (Gabriel et al. 2002.)

Monenlaiset ryhmittelyalgoritmit ja koneoppimiseen perustuvat menetelmät ovat viimeaikoina tuoneet uusia vaihtoehtoja SNP-aineistojen tutkimiseen. Tällaisiin menetelmiin kuuluu mm. Breimanin (2001) kehittämä *Random forest* -algoritmi, joka jaottelee aineistoa luokitteleviin päätöspuihin perustuvilla luokittelijoilla. Algoritmi pyrkii löytämään muuttujien joukosta ne, jotka jakavat yksilöt parhaiten terveisiin ja sairastuneisiin, ja laskee joka muuttujalle sen luokittelukykyyn perustuvan tärkeysindeksin. Genominlaajuisten aineistojen tutkimiseen tämän ajatellaan sopivan hyvin, joten algoritmista on kehitetty SNP-aineistoja varten oma versio *Random Jungle*, jonka julkaisivat Schwarz, König ja Ziegler vuonna 2010.

Tässä tutkielmassa aineisto analysoidaan ensin χ^2-testillä ja logistisella regressioanalyysillä. Näiden testien tulosten perusteella määritellään genomista ne alueet, joilla voidaan epäillä olevan syövän syntymiseen vaikuttavia geenejä. Saatuja tuloksia korjataan hylkäysvirheen takia FDR-menetelmällä. Tämän jälkeen SNP-markkerit ryhmitellään haplotyyppiblokkeihin kytkentäepätasapainoanalyysin avulla ja näin saadut haplotyypit testataan logistisella regressioanalyysillä. Blokeista piirretään myös kuvat *Haploview*-ohjelmalla (Barrett, Fry, Maller & Daly 2005), jolla voidaan tarkastella, miten blokit ovat kromosomeissa jakautuneet. Aineisto analysoidaan vielä *Random forest* -algoritmilla ja algoritmin valitsemat SNP-markkerit testataan vielä χ^2-testillä. Lopuksi tarkastellaan tulosten yhdenmukaisuutta ja pohditaan, onko jokin menetelmä muita parempi syöpäsairauksien ja SNP-markkereiden tutkimiseen.

3.1 χ^2-riippumattomuustesti

Aineisto analysoitiin aluksi käyttäen χ^2-testiä. Kaksiulotteisessa χ^2-testissä verrataan jokaista SNP-markkeria fenotyyppimuuttujaan ja testataan niiden riippumattomuus. Teoreettiset todennäköisyydet kolmen alleelin SNP-markkerin tapauksessa ovat taulukossa 1.

Taulukko 1. SNP-markkerin alleelien teoreettiset todennäköisyydet.

Fenotyyppi	**Alleeli 1**	**Alleeli 2**	**Alleeli 3**	***summa***
0	p_{11}	p_{12}	p_{13}	$\boldsymbol{p_{1+}}$
1	p_{21}	p_{22}	p_{23}	$\boldsymbol{p_{2+}}$
summa	$\boldsymbol{p_{+1}}$	$\boldsymbol{p_{+2}}$	$\boldsymbol{p_{+3}}$	$\boldsymbol{p_{++}}$

Taulukossa 1 p_{ij} on todennäköisyys sille, että havainto on solussa ij. Mikäli henkilö on terve ja hänellä on taulukon 1 SNP-markkerissa alleeli 2, niin $i = 1$ ja $j = 2$ jolloin p_{ij} on siis p_{12}. Merkintä p_{i+} tarkoittaa rivin i yhteenlaskettua todennäköisyyttä ja p_{+j} sarakkeen j todennäköisyyttä, jolloin esimerkiksi sairastuneiden yhteenlaskettu todennäköisyys olisi taulukon 1 tapauksessa p_{2+}. Testattavan hypoteesin H_0 ollessa voimassa oletetaan, että muuttujat ovat riippumattomia. Hypoteesi H_0 on siis

$$H_0: p_{ij} = p_{i+}p_{+j}\,.$$

Hypoteesin H_0 ollessa voimassa todennäköisyyden p_{ij} suurimman uskottavuuden estimaatti $\hat{p}_{ij}$ lasketaan kaavan

$$\hat{p}_{ij} = \hat{p}_{i+}\hat{p}_{+j} = \frac{n_{i+}}{n_{++}}\frac{n_{+j}}{n_{++}}$$

mukaisesti, jossa frekvenssit n_{ij} ovat alleelien havaittuja frekvenssejä kuten taulukossa 2.

Taulukko 2. Alleelien havaitut frekvenssit.

Fenotyyppi	**Alleeli 1**	**Alleeli 2**	**Alleeli 3**	***summa***
0	n_{11}	n_{12}	n_{13}	$\boldsymbol{n_{1+}}$
1	n_{21}	n_{22}	n_{23}	$\boldsymbol{n_{2+}}$
summa	$\boldsymbol{n_{+1}}$	$\boldsymbol{n_{+2}}$	$\boldsymbol{n_{+3}}$	$\boldsymbol{n_{++}}$

Havaittuja frekvenssejä n_{ij} verrataan odotettuihin frekvensseihin e_{ij}, jotka lasketaan kaavan

$$e_{ij} = n_{++}\hat{p}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}$$

mukaisesti. Muuttujien riippumattomuutta kuvaava χ^2-testisuure lasketaan kaavalla

$$\chi_h^2 = \sum_{\mathrm{i}=1}^{\mathrm{r}}\sum_{\mathrm{j}=1}^{\mathrm{s}}\frac{(n_{ij} - e_{ij})^2}{e_{ij}},$$

missä r on rivien ja s sarakkeiden lukumäärä. Hypoteesin H_0 ollessa voimassa testisuure χ_h^2 noudattaa asyptoottisesti χ^2-jakaumaa vapausasteilla $(r - 1)(s - 1)$. Saadusta testisuureesta voidaan laskea p-arvo χ^2-jakauman avulla seuraavasti

$$p = P\left(\chi_{d.f.}^2 \geq \chi_h^2\right),$$

missä d. f. (*degrees of freedom*) merkitsee yllä mainittuja vapausasteita. Fenotyyppimuuttujalla luokkia on kaksi, terveet ja sairastuneet, mutta SNP-markkereilla luokkia voi olla 1–3 alleelien lukumäärästä riippuen. Mikäli saatu p-arvo on pienempi kuin 0,05, voidaan eroja pitää tilastollisesti merkittävinä. (Agresti 1990.)

Testattaessa suurta muuttujajoukkoa pitää ottaa huomioon testattavan hypoteesin hylkäämiseen liittyviä virheitä ja etenkin tyypin 1 virhe, missä testattava hypoteesi H_0 on oikein, mutta testi hylkää sen. Tätä virhettä kutsutaan hylkäysvirheasteeksi (*FDR*, *false discovery rate*) ja sitä voidaan kontrolloida mm. Benjaminin ja Hochbergin (1995) kehittämällä *FDR*-kontrollilla. Merkitään kaikkien valintojen lukumäärää R. Merkitään lisäksi virheellisten valintojen määrän suhdetta kaikkien valintojen määrään $Q = V / R$. Yleensä virheellisten valintojen määrää V ei tiedetä, jolloin myös Q on tuntematon satunnaismuuttuja.

Olkoon *FDR* hylkäysvirheiden osuuden odotusarvo kaikista merkitseviksi valituista p-arvoista eli

$$FDR = E(Q) = P(R > 0)E(V/R\,|R > 0).$$

Olkoot p-arvot järjestetty pienimmästä suurimpaan $p_{(1)}$, $p_{(2)}$,…, $p_{(n)}$, jolloin hylkäysvirheiden kontrollointi voidaan suorittaa valitsemalla ennalta kynnysarvo q. Merkitseväksi valitaan ne p-arvot, jotka eivät ylitä Benjaminin ja Hochbergin kynnysarvoa

$$t_{BH} = \max\left\{p_{(i)} : p_{(i)} \leq \frac{i}{n}q, 0 \leq i \leq n\right\}.$$

Jos kaikki p-arvot ylittävät kynnysarvot t_{BH}, yhtään p-arvoa ei valita merkitseväksi. Benjamini ja Hochberg osoittivat, että jos p-arvot hypoteesin H_0 vallitessa ovat keskenään riippumattomia, *FDR*-valintamenetelmä pyrkii maksimoimaan valittujen p-arvojen luku-määrän niin, että *FDR* on silti annetun kynnysarvon alapuolella.

3.2 Logistinen regressioanalyysi

Logistista regressioanalyysiä pidetään tärkeimpänä luokitteluasteikollisen aineiston mallintamismenetelmänä ja sitä käytetään laajalti erilaisten aineistojen analysointiin. Siinä selitettävä muuttuja Y on binäärinen, eli se voi saada kaksi toisensa poissulkevaa arvoa. Merkitään todennäköisyyttä

$$\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x),$$

missä X on selittävä muuttuja ja x on muuttujan saama arvo.

Yksinkertainen logistinen regressiomalli on muotoa

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}},$$

missä β_0 ja β_1 ovat estimoitavia parametreja. Parametrien suurimman uskottavuuden estimaatteja merkitään $\hat{\beta}_0$ ja $\hat{\beta}_1$.

Olkoon muuttuja Y_i havainnon i fenotyyppi ja olkoon aineiston havaintojen lukumäärä N. Nyt Y_i noudattaa siis binomijakaumaa parametrilla π_i. Tässä aineistossa selittäviä muuttujia ovat SNP-markkerit, joilla voi olla 1–3 luokkaa alleelien lukumäärästä riippuen. Yleensä alleelit luokitellaan kahteen ryhmään sen mukaan, mikä alleeleista on yleisempi (*major allele*) ja mitkä harvinaisempia (*minor allele*). Määritellään muuttuja x_{ij} seuraavasti:

$$x_{ij} = \begin{cases} 1, \text{jos havainnolla } i \text{ on yleisempi alleeli SNP: ssä } j \\ 0, \text{jos havainnolla } i \text{ on harvinainen alleeli SNP: ssä } j \end{cases}.$$

Logistinen regressiomallin mukainen päävaikutusmalli on tällöin muotoa

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{M} x_{ij}\beta_j,$$

missä M on SNP-markkereiden lukumäärä. Testattava hypoteesi on

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_M = 0\,.$$

Tuntemattomat β-parametrit estimoidaan suurimman uskottavuuden menetelmällä. Merkitään muuttujien y_i muodostamaa vektoria $\mathbf{Y}$ ja selittäjien $x_{i1}, x_{i2}, \ldots, x_{im}$ havainnossa i saamien arvojen ja ykkösen muodostamaa vektoria

$$\mathbf{x_i} = (1,\ x_{i1}, x_{i2}, \ldots, x_{im}).$$

Merkitään lisäksi estimoitavien parametrien vektoria

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_m).$$

Nyt logaritminen uskottavuusfunktio on

$$l(\boldsymbol{\beta}; \mathbf{y}) = log[L(\boldsymbol{\beta}; \mathbf{y})] = \sum_{i=1}^{N}\left[y_i\ log\left(\frac{\pi_i}{1-\pi_i}\right) + (1-y_i)log(1-\pi_i)\right],$$

missä

$$\pi_i = \frac{exp\ (\mathbf{x_i'}\boldsymbol{\beta})}{1 + exp\ (\mathbf{x_i'}\boldsymbol{\beta})}\,.$$

Logistisesta regressiomallista seuraa, että logaritminen uskottavuusfunktio voidaan kirjoittaa

$$l(\boldsymbol{\beta};\mathbf{y}) = \sum_{i=1}^{N}\left[y_i \log\left(\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right) + (1-y_i)\log\left(1-\frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right)\right].$$

Suurimman uskottavuuden estimaattien $\widehat{\boldsymbol{\beta}}$ estimointiyhtälöt parametrivektorin $\boldsymbol{\beta}$ suhteen saadaan siis ratkaisemalla derivointilauseke

$$\frac{\partial l(\boldsymbol{\beta};\mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

termeittäin. Määritellään logaritmoitu uskottavuusfunktio parametreille $\boldsymbol{\beta}$

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} y_i\left(\sum_{j=0}^{M} x_{ij}\beta_j\right) - log\left(1+\mathrm{e}^{\sum_{j=0}^{M} x_{ij}\beta_j}\right)$$

ja derivoimalla lauseke $l(\boldsymbol{\beta})$ huomataan, että

$$\frac{\partial}{\partial \beta_j}\sum_{j=0}^{M} x_{ij}\beta_j = x_{ij}\,.$$

Nyt voimme kirjoittaa estimointiyhtälön parametrin β_j suhteen

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{N} y_i\, x_{ij} - \frac{1}{1+\mathrm{e}^{\sum_{j=0}^{M} x_{ij}\beta_j}} \cdot \frac{\partial}{\partial \beta_j}\left(1+\mathrm{e}^{\sum_{j=0}^{M} x_{ij}\beta_j}\right)\\ &= \sum_{i=1}^{N} y_i\, x_{ij} - \frac{1}{1+\mathrm{e}^{\sum_{j=0}^{M} x_{ij}\beta_j}} \cdot \mathrm{e}^{\sum_{j=0}^{M} x_{ij}\beta_j} \cdot x_{ij}\\ &= \sum_{i=1}^{N} y_i\, x_{ij} - \pi_i x_{ij} = 0\,.\end{aligned}$$

Suurimman uskottavuuden estimaattien $\widehat{\boldsymbol{\beta}}$ asymptoottinen kovarianssimatriisin estimaatti $\hat{\sigma}(\widehat{\boldsymbol{\beta}})$ saadaan kääntämällä informaatiomatriisi $\mathbf{I}$, joka koostuu osittaisista logaritmoidun uskottavuusfunktion toisista derivaatoista

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_{k'}} = \frac{\partial}{\partial \beta_{k'}}\sum_{i=1}^{N} y_i\, x_{ij} - \pi_i x_{ij}\,.$$

Mallin sopivuutta testataan Waldin testisuureella

$$Z = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)},$$

missä $\hat{\beta}_j$ on estimoitu parametri β_j ja $\hat{\sigma}(\hat{\beta}_j)$ on estimoitu varianssi parametrille β_j. Hypoteesin H_0 ollessa voimassa testisuure Z noudattaa asymptoottisesti $N(0,1)$ -jakaumaa ja Z^2 noudattaa asymptoottisesti χ^2-jakaumaa. Testisuureelle voidaan laskea p-arvo arvo χ^2-jakauman avulla seuraavasti

$$p = P\left(\chi^2_{d.f.} \geq \chi^2\right),$$

missä vapausasteet d. f. (*degrees of freedom*) ovat (n - 1)(m - 1). (Agresti 1990.) Myös logistisen regressioanalyysin yhteydessä on huomattava hylkäysvirheen mahdollisuus. Virheellisten p-arvojen kontrolloimiseksi käytettiin samaa hylkäysvirheen kontrollointimenetelmää kuin χ^2-testin yhteydessä luvussa 3.1 esitelty Benjaminin ja Hochbergin (1995) *FDR*-menetelmä.

3.3 LD ja haplotyyppiblokit

SNP-markkereiden keskinäistä vuorovaikutusta voidaan kuvata käyttäen alleelien välistä assosiaatiota eli kytkentäepätasapainoa tai LD:tä (*linkage disegilibrium*). LD tarkoittaa lähekkäisten lokusten alleelien esiintymistä yhdessä odotettua useammin. Kahden alleelin, A ja B, lokukselle LD:n perusmitta D määritetään havaittujen ja odotettujen frekvenssien erotuksena. Kytkentäepätasapainon määritelmästä ja laskutavoista julkaistiin artikkeli jo vuonna 1960, jolloin kirjoittajina olivat Lewontin ja Kojima.

Merkitään kahden SNP-markkerin alleelien A ja B muodostamien haplotyyppien A-A, A-B, B-A ja B-B todennäköisyyksiä p_{11}, p_{12}, p_{21} ja p_{22}. Mikäli siis markkerin SNP1 alleelit ovat A ja G ja markkerin SNP2 alleelit C ja T, ovat haplotyypit muotoa AC, AT, GC ja GT. Alleelin AC todennäköisyyttä vastaa siis merkintä p_{11}, alleelin AT todennäköisyyttä p_{12} jne. Alleelien frekvenssit voidaan kirjoittaa ristiintaulukkoon, kuten taulukossa 3.

Taulukko 3. SNP1- ja SNP2 -markkereiden alleelien frekvenssitaulukko.

	A	**G**	**summa**
C	n_{11}	n_{12}	n_{1+}
T	n_{21}	n_{22}	n_{2+}
summa	n_{+1}	n_{+2}	n_{++}

Merkitään lisäksi alleelien summafrekvenssejä muuttujilla n_{1+}, n_{2+}, n_{+1} ja n_{+2}. Tällöin lokusten välinen kytkentäepätasapainomitta määritetään seuraavasti:

$$D = n_{11}n_{22} - n_{12}n_{21}\ .$$

Yleensä kytkentäepätasapainomitan *D* lisäksi ilmoitetaan skaalattu kytkentäepätasapainomitta *D'* sekä korrelaatiokertoimen neliö r^2, jotka lasketaan seuraavien kaavojen mukaan:

$$D' = \frac{D}{D_{max}}, \text{missä } D_{max} = \begin{cases} \min(n_{1+}n_{+2}, n_{+1}n_{2+}) \text{ kun } D \geq 0 \\ \min(n_{1+}n_{+1}, n_{+2}n_{2+}) \text{ kun } D < 0 \end{cases}$$

$$r^2 = \frac{D^2}{n_{1+}n_{2+}n_{+1}n_{+1}} .$$

Sanotaan, että SNP on kytkentäepätasapainossa, mikäli *D* eroaa merkittävästi nollasta. (Ollikainen & Uimari 2004.)

SNP-parille lasketaan lisäksi *LOD*-arvo (*logarithm of odds*), joka kuvaa kytkentäepätasapainon tilastollista merkitsevyyttä. Olkoon yksilöiden määrä otoksessa *n* ja olkoot n_d ja n_c sairastuneiden ja terveiden lukumäärät otoksessa. Kun SNP-markkereiden alleeleja merkitään kirjaimin AA, AB ja BB, olkoot n_{AA}, n_{AB} ja n_{BB} niiden yksilöiden lukumäärät, joilla on kyseinen alleeli. Tässä alleelit AB ja BA ajatellaan samoiksi. Jokaiselle otoksen genotyypille lasketaan myös sairastuneiden ja terveiden frekvenssit, joita merkitään $n_{AA,D}$, $n_{AB,D}$, $n_{BB,D}$, $n_{AA,C}$, $n_{AB,C}$ ja $n_{BB,C}$. Olkoon ψ sairauden esiintyvyys populaatiossa. Merkitään genotyypin g periytyvyyttä populaatiossa ψ_g, missä g voi olla AA, AB tai BB. Olkoon *y* muuttuja, joka ilmaisee onko yksilö sairas tai terve niin, että *y* saa arvon 0, kun yksilö on terve, ja 1, kun yksilö on sairastunut. Taulukossa 4 on esimerkki, miten frekvenssit jakautuvat alleelien kesken.

Taulukko 4. Kahden SNP-markkerin alleelien sairastuneiden ja terveiden frekvenssit.

	SNP1		
SNP2	*A*	*B*	**Fenotyyppi**
A	$n_{AA,D}$	$n_{AB,D}$	*sairastunut*
	$n_{AA,C}$	$n_{AB,C}$	*terve*
B	$n_{AB,D}$	$n_{BB,D}$	*sairastunut*
	$n_{AB,C}$	$n_{BB,C}$	*terve*

Yksilön sairastuvuuden todennäköisyysfunktio voidaan kirjoittaa

$$l(y|g) = (\psi_g)^y (1 - \psi_g)^{1-y}.$$

Tutkittavan SNP-markkerin alleelit huomioiden uskottavuusfunktio *L* satunnaisten *n* yksilön sairastuvuudelle on muotoa

$$L = \prod_{i=1}^{n} l_i(y|g).$$

Hypoteesin H_0 ollessa voimassa oletetaan, että SNP-markkerit eivät ole sairautta aiheuttavia.

Silloin kaikkien alleelien periytyvyys on yhtä suuri kuin ψ, jolloin uskottavuus *L* on

$$L_0 = \psi^{n_D}(1-\psi)^{n_C}.$$

Vaihtoehtoisen hypoteesin H_1 ollessa voimassa SNP-markkereilla epäillään olevan vaikutusta sairastuvuuteen, jolloin uskottavuus *L* on muotoa:

$$L_1 = {\psi_{AA}}^{n_{AA,D}}(1-\psi_{AA})^{n_{AA,C}}{\psi_{AB}}^{n_{AB,D}}(1-\psi_{AB})^{n_{AB,C}}{\psi_{BB}}^{n_{BB,D}}(1-\psi_{BB})^{n_{BB,C}}$$

Nyt *LOD* voidaan kirjoittaa muotoon

$$LOD = \log_{10}\left[\frac{\hat{L}_1}{\hat{L}_0}\right],$$

missä $\hat{L}_0$ ja $\hat{L}_1$ on laskettu eri ψ:n arvojen suurimman uskottavuuden estimaattien avulla. (Deng & Recker, 2001.)

Blokki muodostetaan, mikäli vertailtavista SNP-markkereista 95 % voidaan luokitella voimakkaan LD:n alaisiksi. SNP-parin sanotaan olevan voimakkaan LD:n alaisia, mikäli skaalatun kytkentäepätasapainomitan *D'* yksipuolinen 95 % luottamustaso on suurempi kuin 0,98. Määritellään aineiston perusteella lasketun skaalatun kytkentäepätasapainomitan uskottavuusfunktio niin, että kun $k = 1, 2, \ldots, 100$, niin

$$l(k) = \mathrm{P}(data||D'| = 0{,}01 \cdot k).$$

Olkoon luottamustason alaraja

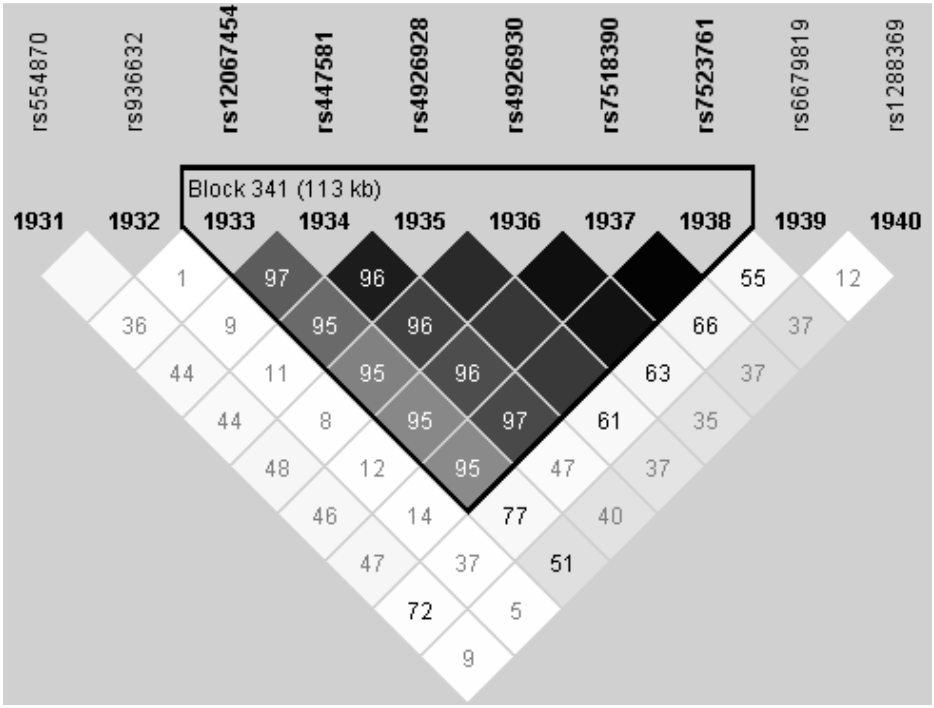
$$C_L = \frac{\sum_{i=1}^{k-1} l(i)}{\sum_{i=1}^{100} l(i)} \leq 0{,}05$$

ja yläraja

$$C_U = \frac{\sum_{i=k+1}^{100} l(i)}{\sum_{i=1}^{100} l(i)} \leq 0{,}05\,.$$

Siis jos $C_L \geq 0{,}7$ ja $C_U \geq 0{,}9$ sanotaan SNP-parin olevan voimakkaan LD:n alaisia ja kuuluvan samaan blokkiin. (Gabriel et al. 2002.)

Haplotyyppikuvissa värityksen perusteena ovat *D'*- ja LOD-mittaukset siten, että mitä tummempi väri, sitä korkeampi *D'*-arvo ja tilastollisesti merkittävämpi LOD-tulos. Kuvassa 4 on esimerkki haplotyyppikuvasta ja sen värityksistä. Ruudun sisällä oleva numero on kytkentäepätasapainomitan *D* arvo. Mikäli ruutu on tyhjä, *D* on pienempi kuin 1,0. (Barrett et al. 2005.)



Kuva 4. Haplotyyppiblokki 341 kromosomista 1.

Haplotyyppiblokkien merkitsevyyttä voidaan myös testata χ^2-testillä sekä logistisella regressioanalyysillä. Blokin frekvenssidata voidaan järjestää 2 x *k* ristiintaulukkoon, missä *k* on haplotyypin emästen lukumäärä. Testattavan hypoteesin H_0 ollessa voimassa oletetaan, että haplotyypin frekvenssit sairastuneiden ja verrokkien osalta ovat samansuuruisia. Testaamiseen käytettävä χ^2 -testisuure määritellään

$$\chi^2_{HT} = 2n \sum_{l=1}^{k} \frac{\left(\hat{P}_{Dl} - \hat{P}_{Cl}\right)^2}{\hat{P}_{Dl} + \hat{P}_{Cl}},$$

missä $\hat{P}_{Dl}$ ja $\hat{P}_{Cl}$ ovat *l.* haplotyypin havaitut sairastuneiden ja verrokkien frekvenssit. Testattavan hypoteesin H_0 ollessa voimassa, χ^2_{HT} noudattaa asymptoottisesti χ^2_{k-1} jakaumaa. (Akey, Jin & Xiong 2001.)

Määritellään S_{ij} seuraavasti:

$$S_{ij} = \begin{cases} 1, kun\ havainnolla\ i\ on\ haplotyyppi\ j \\ 0, kun\ havainnolla\ i\ ei\ ole\ haplotyyppiä\ j \end{cases}.$$

Oletetaan, että $Y_i \sim Bin(1, \pi_i)$, jolloin logistinen regressiomalli on muotoa

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=1}^{h} S_{ij}\beta_j\,,$$

missä *h* on haplotyyppien lukumäärä.

Mallin sopivuutta voidaan testata Waldin testisuureella

$$Z = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)},$$

missä $\hat{\beta}_j$ on estimoitu parametri β_j ja $\hat{\sigma}(\hat{\beta}_j)$ on estimoitu varianssi parametrille β_j. Kaavat suurimman uskottavuuden estimaattien määrittämiseen on esitetty luvussa 3.2 logistisen regressioanalyysin yhteydessä. Testisuure Z noudattaa asymptoottisesti $N(0,1)$ -jakaumaa. (Wason & Dudbridge 2010.)

3.4 Random forest -algoritmi

Random forest -algoritmi on luokittelumenetelmä, joka koostuu kokoelmasta päätöspuiden tapaisista luokittelijoista. Luokittelupuut ryhmittelevät aineistoa sen piirteiden mukaan osajoukkoihin, kunnes luokittelu on täydellinen eli kussakin osajoukossa on vain samaan luokkaan kuuluvia havaintoja. RF-algoritmissa puita muodostetaan satunnaisesti bootstrap-otosten avulla suuri määrä ja muodostettujen puiden avulla voidaan selvittää muuttujalle sopivin luokka. Aineisto jaetaan puiden muodostamista ja niiden luokittelun testaamista varten opetus- ja testausaineistoiksi.

Breimanin vuonna 2001 kehittämän algoritmin kulku voidaan yksinkertaistaen kuvata näin: Otetaan n_{tree} suuruinen bootstrap-otos B^* alkuperäisestä aineistosta. Jokaisesta bootstrap-otoksesta kasvatetaan luokittelupuu T_b, jonka jokaisessa solmukohdassa (*node*) otetaan m_{try} suuruinen ennustajamuuttujien satunnaisotos. Tämän otoksen perusteella valitaan parhain jako, jonka mukaan havainnot jaetaan ryhmiin. Näin jatketaan, kunnes jako havaintojen kesken on täydellinen. Jokaiselle ennustajalle lasketaan solmukohdassa tärkeysindeksi, joka kuvaa ko. ennustajan kykyä luokitella muuttujat. Bootstrap-otoksen ulkopuolelle jääneitä havaintoja kutsutaan *oob*-aineistoksi (*out of bag*), jonka avulla voidaan arvioida luokitteluvirhettä ja määritetään muuttujien tärkeysindeksi. Kasvatettavien puiden lukumäärä n_{tree} valitaan niin, että luokitteluvirhe olisi mahdollisimman pieni.

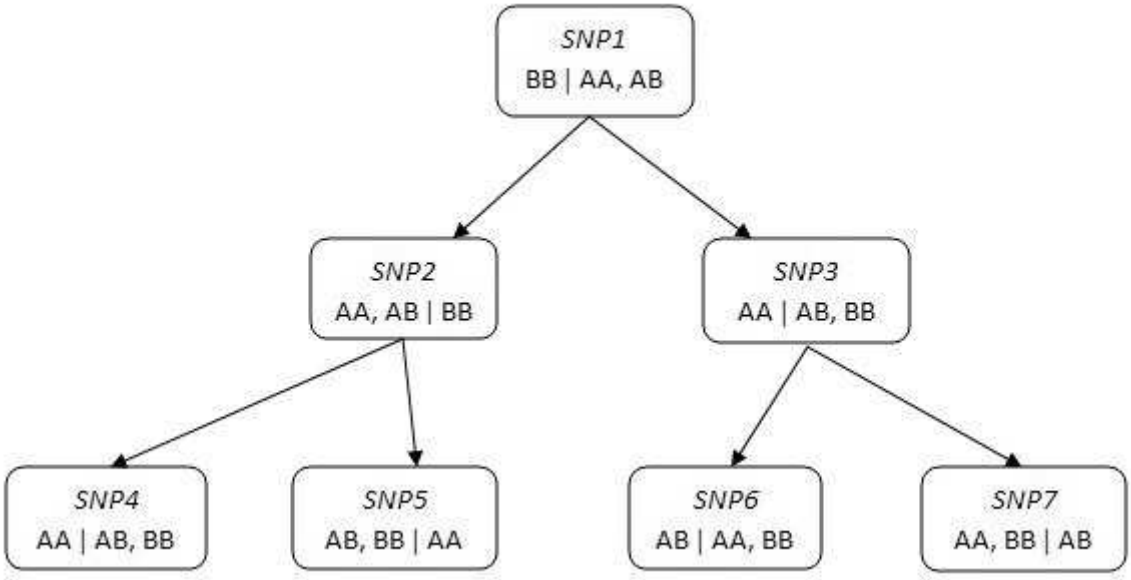
RF-algoritmille on olemassa myös versio regressioanalyysiä varten, mutta luokittelualgoritmi on tämän tutkielman aineistoon sopivampi. Luokitteleva RF-algoritmi pystyy käsittelemään nopeasti moniluokkaisia aineistoja, joten sitä pidetään sopivana SNP-aineistojen luokitteluun. Tarjolla olevista ohjelmistoista päädyimme käyttämään erityisesti SNP-aineistoja varten räätälöityä *Random Jungle* -ohjelmistoa. (Schwarz, König & Ziegler 2010.)

3.4.1 Luokittelupuut

Luokittelupuihin perustuvien menetelmien tarkoituksena on jakaa aineistoa ominaisuuksiensa perusteella osajoukkoihin, kunnes jokaisessa osajoukossa on vain samankaltaisia havaintoja. Luokittelupuut muodostuvat solmukohdasta (*node*) ja oksista (*branch*) siten, että solmukohtiin valitaan havaintoja hyvin luokitteleva muuttuja, ja muodostetut osajoukot ovat solmukohdasta haarautuvia oksia. Puulle voidaan myös määrätä juuri (*root*) eli algoritmin aloittava muuttuja. Puun päättävää solmukohtaa sanotaan päätesolmuksi (*terminal node*). Tämän tutkielman aineistossa havainnot pyritään jakamaan kahteen luokkaan siten, että toisessa olisi pelkkiä syöpään sairastuneita ja toisessa terveitä yksilöitä. SNP-markkereiden luokittelua varten markkereiden alleelit yleensä jaetaan kahteen ryhmään sen mukaan, miten ne jakavat

aineiston havainnot parhaiten terveisiin ja sairastuneisiin. (Hastie, Tibshirani & Friedman 2001.)

Kuvassa 5 on esimerkki luokittelupuun osasta. Siinä ensimmäisen SNP-markkerin alleelit jaetaan ryhmiin niin, että alleelin *BB* havainnot menevät vasemmanpuoleista oksaa ja oikealle menevät alleelien *AA* ja *AB* havainnot. Olkoon alleelin BB havaintojen sairastuneiden frekvenssi suurempi kuin alleeleilla *AA* ja *AB*. Seuraavien SNP-markkereiden kohdalla jako pyritään tekemään samoin niin, että vasemmalle menevissä osajoukoissa sairastuneiden frekvenssi olisi suurempi kuin oikealle. Lopulta puun aivan vasemmanpuoleisessa lehdessä tulisi olla pelkkiä sairastuneita ja oikeanpuoleisessa vain terveitä havaintoja.



Kuva 5. SNP-markkereita luokittelupuussa.

SNP-markkerin alleelit muunnetaan solmukohdissa binäärisiksi sen mukaan, millä alleelijaolla SNP luokittelee havainnot parhaiten. Alleeli saa binäärimerkinnän 1, kun ko. alleelin omistavat havainnot jakautuvat oikealle, ja 0, kun alleelin havainnot jakautuvat vasemmalle. Esimerkiksi kuvassa 5 SNP1-markkerin alleelien binäärimuoto olisi (1, 1, 0) ja SNP2-markkerin (0, 0, 1). Puun juuri eli ensimmäinen luokittelumuuttuja voidaan määrätä tai antaa algoritmin valita se satunnaisesti. Jokaisessa solmukohdassa algoritmi sovittaa muuttujia ennalta määrätyn määrän, jota merkitään parametrilla m_{try}. Algoritmi pyrkii löytämään jokaiseen solmukohtaan sellaisen muuttujan, joka siinä tilanteessa jakaa havainnot parhaiten.

Tarkastellaan aineistoa, jossa ennustettavia havaintoja merkitään y_i (i = 1, 2, …, N) ja jossa luokkamuuttuja saa arvot 1, 2, …, K. Olkoon puun solmukohdassa m, jossa on N_m havainnon suuruinen osajoukko R_m, luokan k havaintojen osuus

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k),$$

missä I on indikaattorifunktio, joka saa arvon 1 kun y_i kuuluu luokkaan k ja 0 kun y_i ei kuulu luokkaan k.

Havainnot solmukohdassa *m* luokitellaan solmun enemmistöluokkaan seuraavasti:

$$k(m) = \arg\max_k \hat{p}_{mk} .$$

Solmulle *m* voidaan määritellä luokitteluvirhe seuraavasti:

$$\frac{1}{N_m} \sum_{i \in R_m} I\big(y_i \neq k(m)\big) = 1 - \hat{p}_{mk(m)} .$$

Puiden luokittelun arvioimista varten aineisto jaetaan opetus- ja testausaineistoiksi. Puut muodostetaan käyttämällä opetusaineistoa ja valmiiden puiden luokittelukykyä arvioidaan testausaineistolla. (Hastie et al. 2001.)

3.4.2 Bootstrap-otokset ja *oob*-aineisto

Bootstrap-menetelmä on otosmenetelmä, jossa alkuperäisestä aineistosta valitaan satunnaisesti havaintoja otoksiin esimerkiksi estimaattien tarkkuuden arviointia varten. Menetelmän kehitti B. Efron vuonna 1979. Bootstrap-otoksia voidaan tehdä joko palauttaen tai palauttamatta, jolloin aineiston havainto voi tulla valituksi otokseen vain kerran. Oletetaan, että opetusaineisto muotoa $T = \{(y_n, x_n),\ = 1, \dots, N\}$. Olkoon aineistosta muodostettavien otosten lukumäärä *B*, ja kutsutaan otoksia bootstrap-otoksiksi. Todennäköisyys, että havainto *i* kuuluu bootstrap-otokseen *b* voidaan kirjoittaa muotoon

$$p = 1 - \left(1 - \frac{1}{N}\right)^N .$$

Breimanin 2001 kehittämässä *Random forest* -algoritmissa bootstrap-otokset ovat osa *bagging*-menetelmää, jolla pyritään vähentämään estimoidun ennustusfunktion varianssia. Bagging-menetelmästä on käytetty myös termiä bootstrap-kokoelma (*bootstrap aggregation*). Se kuuluu koneoppimisen algoritmeihin ja vaikuttaa toimivan erityisen hyvin luokittelupuiden kaltaisten menetelmien yhteydessä: *RF* -algoritmissa luokittelupuiden joukko äänestää muuttujan luokasta, joten bagging-menetelmä pienentää luokitteluvirhettä.

Oletetaan, että yllä määritetylle opetusaineistolle *T* on määritelty ennuste $\hat{f}(x)$ arvolle *x*. Bagging-menetelmässä lasketaan sama ennuste kokoelmalle bootstrap-otoksia ja otetaan näistä bootstrap-ennusteista keskiarvo. Merkitään bootstrap-otoksesta T^{*b} laskettua ennustajaa $\hat{f}^{*b}(x)$. Bagging-estimaatti voidaan kirjoittaa muotoon

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x) .$$

Koska bagging-menetelmällä muodostetut puut ovat identtisesti jakautuneita, puiden keskiarvon odotusarvo on kaikille sama. (Breiman 1996.) *RF* -menetelmässä käytetään *out-of-bag* -otoksia, jotka ovat bootstrap-otosten ulkopuolelle jääneitä havaintoja. Niitä käytetään luokitteluvirheen estimoinnissa ja tärkeysindeksin laskemisessa, josta on tarkempi selvitys seuraavassa luvussa.

3.4.3 Tärkeysindeksi

Tärkeysindeksi (*importance index*) kuvaa luokittelumuuttujan kykyä erotella havainnot luokkiin. Sen perusteella SNP-markkereiden joukkoa pienennetään analyysikierrosten välissä siten, että pienen indeksiarvon saaneet markkerit jätetään seuraavien analyysien ulkopuolelle. Kaikissa puissa jokaisen solmun kohdalla jakautumistarkkuuden parantuminen kerryttää luokittelumuuttujan tärkeysindeksiä. Kaikilta puilta luokittelumuuttujalle lasketuista tärkeysindeksiarvoista otetaan lopuksi keskiarvo. Koska jokaisessa solmukohdassa algoritmi sovittaa ennalta määrätyn m_{try} parametrin verran luokittelumuuttujia, on tärkeää asettaa parametrin m_{try} arvo tarpeeksi korkeaksi, jotta kaikilla muuttujilla tärkeysindeksi kertyisi tasapuolisesti.

Oletetaan, että opetusaineisto on muotoa $T = \{(y_n, x_n),\ n = 1, \dots, N\}$, ja että on olemassa menetelmä ennustajan *F(y,T)* muodostamiseen opetusaineiston pohjalta. Olkoon tulosmuuttuja *y* luokitteluasteikollinen muuttuja. Opetusaineistojen sarja $T_{B,i}$, …, $T_{B,K}$ generoidaan bootstrap-otoksilla opetusaineistosta *T*. Ennustajat *K* muodostetaan niin, että *k*. ennustaja $F(x, T_{k,B})$ perustuu *k*. bootstrap-otokseen opetusaineistosta *T*. Ennustajat antavat äänen yksilölle sopivimmalle luokalle.

Olkoon X_i yksilön *i* ennustajamuuttujien vektori, y_i yksilön oikea luokka, $V_j(X_i)$ puun *j* antama ääni ja t_{ij} indikaattorimuuttuja, joka saa arvon 1 kun yksilö *i* on puun *j oob*-havainto ja 0 muulloin. Olkoon

$$X^{(A,j)} = \left(X_1^{(A,j)}, \dots, X_N^{(A,j)}\right)$$

ennustajamuuttujien vektori muuttujan *A* satunnaisesti puulle *j* permutoitujen *oob*-havaintojen kanssa, ja olkoon vektoreiden $X^{(A,j)}$ kaikkien puiden kokoelma $X^{(A)}$, missä N on otoksen yksilöiden lukumäärä. Olkoon vielä *1(C)* indikaattorifunktio, joka saa arvon 1 kun ehto C on tosi ja 0 muulloin. Nyt tärkeysindeksi on keskiarvo kaikilta puilta ja voidaan kirjoittaa muotoon

$$I_T(A) = \frac{1}{T}\sum_{j=1}^{T}\frac{1}{N_j}\sum_{i=1}^{N}\left[1\left(V_j(X_i) = y_i\right) - 1\left(V_j\left(X_i^{(A,j)}\right) = y_i\right)\right] t_{ij}\ ,$$

missä N_j on puun *j oob*-havaintojen lukumäärä ja *T* puiden kokonaislukumäärä. Standardoitu tärkeysindeksi saadaan jakamalla puun välisistä variansseista saadulla keskihajonnalla:

$$Z_T(A) = {I_T(A)}\Big/{\sqrt{\sigma_T^2(A)/T}}\,.$$

Varianssi σ_T^2 on tärkeysindeksin I_T puukohtainen varianssi. Hypoteesin H_0 ollessa voimassa standardoitu tärkeysindeksi Z_T noudattaa asymptoottisesti *N(0,1)*-jakaumaa. (Lunetta, Hayward, Segal & van Eerdewegh 2004.)

4 Aineiston analyysi

Tämän tutkielman tarkoituksena on löytää yli 91 000 SNP-markkerin joukosta sellaiset, joilla voi olla vaikutusta rintasyövän kehittymiselle. Näin voidaan paikantaa genomista sellaiset alueet, joilla mahdolliset geneettiset mutaatiot altistavat syövälle. Yksi SNP ei itsessään aiheuta tai ehkäise syöpää, mutta monen yksittäisen SNP-muutoksen yhteisvaikutus voi olla merkittävä. SNP-joukko olisi tarkoitus rajata pienemmäksi niin, että jatkotutkimukset olisivat helpommin tehtävissä.

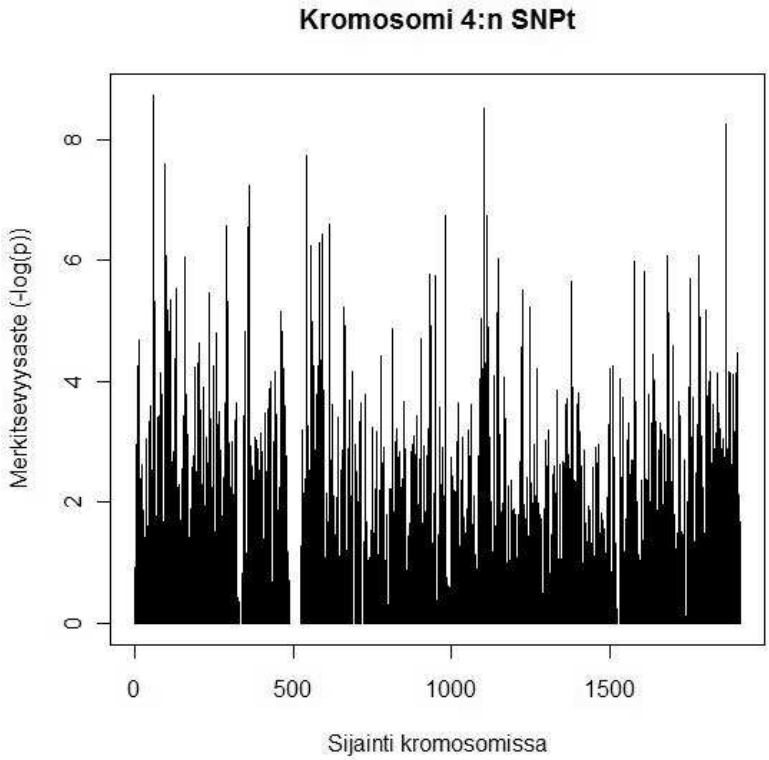
Aineiston SNP-markkerit analysoitiin ensin yksi kerrallaan käyttäen sekä χ^2 -testiä että logistista regressioanalyysiä. Testien tuloksia vertailtiin ja tulosten avulla piirrettiin kromosomikohtaiset p-arvokuvaajat, joista syövän kannalta merkitsevien markkerialueiden havainnointi käy helposti. Nämä analyysit tehtiin *plink*-ohjelmalla (Purcell et al. 2007). Tällaiset analyysit eivät ota huomioon SNP-markkereiden riippuvuutta toisistaan, joten tuloksia on syytä tarkastella kriittisesti ns. väärien positiivisten tulosten takia. Hylkäysvirhettä arvioitiin Benjaminin ja Hochbergin (1995) kehittämällä *FDR*-kontrollilla.

SNP-markkereiden välinen riippuvuus aiheuttaa monia ongelmia yksinkertaisille assosiaatiotesteille. Tämän takia SNP-markkerit ryhmiteltiin blokkeihin käyttäen kytkentäepätasapainomääritelmää. Saman blokin SNP-markkereiden alleelien eri yhdistelmiä kutsutaan haplotyypeiksi, joiden vaikutusta syövän syntyyn testattiin logistisella regressioanalyysillä. Ryhmittely blokkeihin ja haplotyyppien analyysi tehtiin *plink*-ohjelmalla (Purcell et al. 2007). *Haploview*-ohjelmalla blokeista piirrettiin kuvat, joista voidaan tarkastella blokkien sijoittumista kromosomeihin (Barrett et al. 2005).

Aineiston analysointiin käytettiin Breimanin 2001 kehittämää *Random forest* -algoritmia, sillä sen oletetaan sopivan hyvin genominlaajuisten SNP-aineistojen analysointiin. Algoritmi etsii SNP-markkereiden joukosta sellaiset, jotka parhaiten jakavat yksilöt sairastuneiden ja kontrollitapausten kesken. Jokaiselle markkerille algoritmi laskee tärkeysindeksin, joka kuvaa sen kykyä erotella terveet ja sairastuneet. Tärkeysindeksin perusteella valitaan sellaiset markkerit, joilla voidaan olettaa olevan vaikutusta syövän syntymiselle. Algoritmin valitsemille SNP-markkereille tehtiin vielä χ^2 -testi, jotta varmistuttaisiin algoritmin kyvystä löytää merkitsevät markkerit. Analysointiin käytettiin *Random Jungle* -ohjelmaa (Schwarz et al. 2010).

4.1 Assosiaatiotestaus ja logistinen regressio

χ^2 -testin mukaan koko aineistossa on 1 282 SNP-markkeria merkitsevyysasteella 0,01 ja 5 516 merkitsevyysasteella 0,05. Vastaavat lukumäärät logistisen regressioanalyysin perusteella ovat 1 123 ja 5 255. Logistisen regressiomallin 5 255 merkitsevimmästä SNP-markkerista 5 134 oli myös χ^2-testin merkittävimpien joukossa. Logistisen regressioanalyysin tulosten perusteella joka kromosomille piirrettiin kuvan 6 kaltainen p-arvokuvaaja, jossa vaaka-akselilla on SNP-markkerin sijainti kromosomissa ja pystyakselilla logaritmoidun p-arvon vastaluku. Näistä kuvaajista voi silmämääräisesti arvioida, missä päin kromosomia on sairauden syntymiseen vaikuttavia SNP-markkereita.



Kuva 6. Kromosomi 4:n p-arvokuvaaja.

Taulukossa 5 on logistisen regressioanalyysin perusteella 15 merkitsevintä SNP-markkeria. Taulukossa on ilmoitettu SNP-markkerin kromosomi, lokus, markkerille estimoitu parametrin β_j arvo ja t-testin p-arvo. Mitä pienempi p-arvo on, sitä suuremmalla varmuudella voidaan sanoa, että $\beta_j \neq 0$ ja markkerilla on vaikutusta yksilön fenotyyppiin. Positiivinen β_j arvo merkitsee, että alleelin vaihtuminen yleisemmästä harvinaisempaan nostaa yksilön riskiä kuulua sairastuneiden ryhmään. Samantapainen taulukko saatiin myös χ^2-testin tuloksena, jonka 15 merkitsevintä SNP-markkeria on liitteenä (Liite 1).

Taulukko 5. Logistisen regressioanalyysin perusteella merkitsevimmät SNP:t.

Kromosomi	SNP	Lokus	β_j	p-arvo
17	rs8066558	27296992	3,695	1,91E-15
10	rs7905923	109663117	1,227	3,72E-06
19	rs4807425	3180224	0,9267	5,74E-06
17	rs319749	28596020	1,025	2,00E-05
10	rs7915527	109721574	1,04	2,27E-05
18	rs965174	54995424	0,7448	2,80E-05
10	rs4255475	48144906	1,434	3,75E-05
11	rs541207	63881718	-0,7651	4,34E-05
2	rs6738882	37890802	0,9423	4,42E-05
9	rs7046415	80670516	-0,8467	5,18E-05
2	rs10193128	233695966	0,7111	6,20E-05
21	rs2834950	35803265	0,7674	8,16E-05
9	rs12336488	9218107	0,7145	8,95E-05
5	rs6450641	28588040	-0,7245	0,000111
14	rs85425	58454706	1,318	0,000113

Kromosomikohtainen yhteenveto testien tuloksista on taulukossa 6. Siinä on listattu jokaisen kromosomin osalta testikohtaisesti merkitsevyystasojen 0,01 ja 0,05 alapuolelle jäävien SNP-markkereiden lukumäärät sekä merkitsevyystason 0,05 alapuolelle jäävien SNP-markkereiden osuus kaikista kromosomin markkereista. Alimmalla rivillä on merkitsevien SNP-markkereiden summat sekä keskiarvo merkitsevien SNP-markkereiden prosenttiosuuksista.

Taulukko 6. Kromosomikohtaisia χ^2-testin ja logistisen regressioanalyysin tuloksia.

	χ^2-testi			Logistinen regressioanalyysi		
Kromosomi	**0,01**	**0,05**	**0,05-%**	**0,01**	**0,05**	**0,05-%**
1	84	386	5,687	69	355	5,197
2	84	379	5,134	71	363	4,905
3	71	335	5,903	60	326	5,733
4	75	312	6,637	69	299	6,346
5	103	352	6,156	88	332	5,796
6	72	379	6,519	63	360	6,183
7	69	276	5,693	60	263	5,363
8	59	317	5,608	49	310	5,473
9	75	273	6,421	72	257	6,022
10	72	260	5,912	63	249	5,648
11	74	297	6,928	66	283	6,575
12	54	278	6,546	50	255	5,968
13	29	154	5,147	27	143	4,768
14	30	163	5,699	27	155	5,399
15	43	161	5,752	34	150	5,293
16	35	166	6,234	35	155	5,764
17	34	157	5,891	30	152	5,651
18	40	164	6,094	37	155	5,739
19	28	119	6,560	24	118	6,417
20	37	150	6,266	35	145	6,034
21	31	112	6,888	27	106	6,487
22	24	100	6,349	23	99	6,219
23	59	221	6,619	43	222	6,505
yht.	**1282**	**5511**	**6,054**	**1122**	**5252**	**5,729**

Molempien testien perusteella suhteellisesti eniten merkitseviä SNP-markkereita on kromosomissa 11, sillä sen merkitsevien markkereiden osuus kaikista markkereista on suurin (χ^2-testillä 6,9 %, logistisen regressioanalyysin perusteella 6,6 %). Muita keskimääräistä enemmän merkitseviä markkereita sisältäviä kromosomeja ovat 4, 19, 21 ja 23. Pienimpiä merkitsevien SNP-markkereiden prosenttiosuuksia on kromosomeilla 1, 2 ja 13, tosin kromosomeilla 1 ja 2 on huomattavasti enemmän SNP-markkereita kuin muilla kromosomeilla (kuva 3, luku 2.3).

Taulukko 7. Assosiaatiotestien korjattuja p-arvoja.

Logistinen regressio				χ^2-testi			
Kromosomi	**SNP**	**p-arvo**	**BH-arvo**	**Kromosomi**	**SNP**	**p-arvo**	**BH-arvo**
17	rs8066558	1,91E-15	1,75E-10	17	rs8066558	4,17E-38	3,80E-33
10	rs7905923	3,72E-06	0,1705	10	rs7905923	1,70E-06	0,04462
19	rs4807425	5,74E-06	0,1754	1	rs12123980	1,80E-06	0,04462
17	rs319749	2,00E-05	0,416	19	rs4807425	1,96E-06	0,04462
10	rs7915527	2,27E-05	0,416	17	rs319749	5,44E-06	0,09535

Testien tuloksille tehtiin myös hylkäysvirheen korjaus Benjaminin ja Hochbergin (1995) kehittämällä *FDR*-kontrollimenetelmällä. Molempien testien viisi merkitsevintä SNP-markkeria sekä alkuperäisten että korjattujen p-arvojen (BH-arvo) kanssa ovat taulukossa 7. Siitä huomataan, että virhekontrollin jälkeen merkitseviä SNP-markkereita logistisen regressioanalyysin mukaan olisi vain yksi. Myös χ^2-testin tuloksissa alle 0,01 merkitsevyystason SNP-markkerit vähenivät seitsemään, joten testit eivät anna luotettavia tuloksia SNP-markkereiden riippuvuuden takia. Tämän takia aineiston analysointiin valittiin lisäksi menetelmiä, jossa SNP-markkereiden riippuvuus otetaan huomioon.

4.2 LD ja haplotyyppiblokit

SNP-markkerit jaettiin kytkentäepätasapainomääritelmän mukaisesti ryhmiin eli blokkeihin käyttäen *plink*-ohjelmistoa. Tulosteena saadaan taulukon 8 mukainen lista, missä blokit on järjestetty kromosomin ja sijainnin mukaan. BP1 on blokin aloituskohta, BP2 lopetuskohta ja KB blokin pituus tuhansina emäksinä. Lisäksi ilmoitetaan blokissa sijaitsevien SNP-markkereiden lukumäärä ja niiden nimet.

Taulukko 8. Kromosomin 1 ensimmäiset 10 blokkia.

Kromo	BP1	BP2	KB	SNP lkm	SNPt
1	1020428	1038818	18.391	2	rs6687776|rs4970405
1	1084601	1089205	4.605	2	rs4970362|rs9660710
1	1696020	1799950	103.931	4	rs7531583|rs742359|rs6681938|rs7525092
1	2016609	2023116	6.508	2	rs884080|rs908742
1	2136826	2146222	9.397	2	rs7512482|rs2460000
1	2170384	2194615	24.232	3	rs260513|rs7547453|rs7553178
1	2273756	2289487	15.732	4	rs2840528|rs903914|rs2840542|rs7545940
1	2310562	2328934	18.373	3	rs4648633|rs3001336|rs2494428
1	2362949	2369108	6.16	2	rs6659405|rs11581548
1	2480088	2497275	17.188	2	rs2234167|rs4310388

Koko aineistossa blokkeja määritettiin 16 277. Eniten blokkeja oli kromosomissa 2 (1 285) ja vähiten kromosomissa 19 (282). Keskimäärin blokkeja on noin 700 yhdessä kromosomissa. Yhdessä blokissa oli keskimäärin 2–3 SNP-markkeria ja enimmillään niitä oli kromosomissa 5 olleessa blokissa 14 kpl, josta on kuva liitteessä 2.

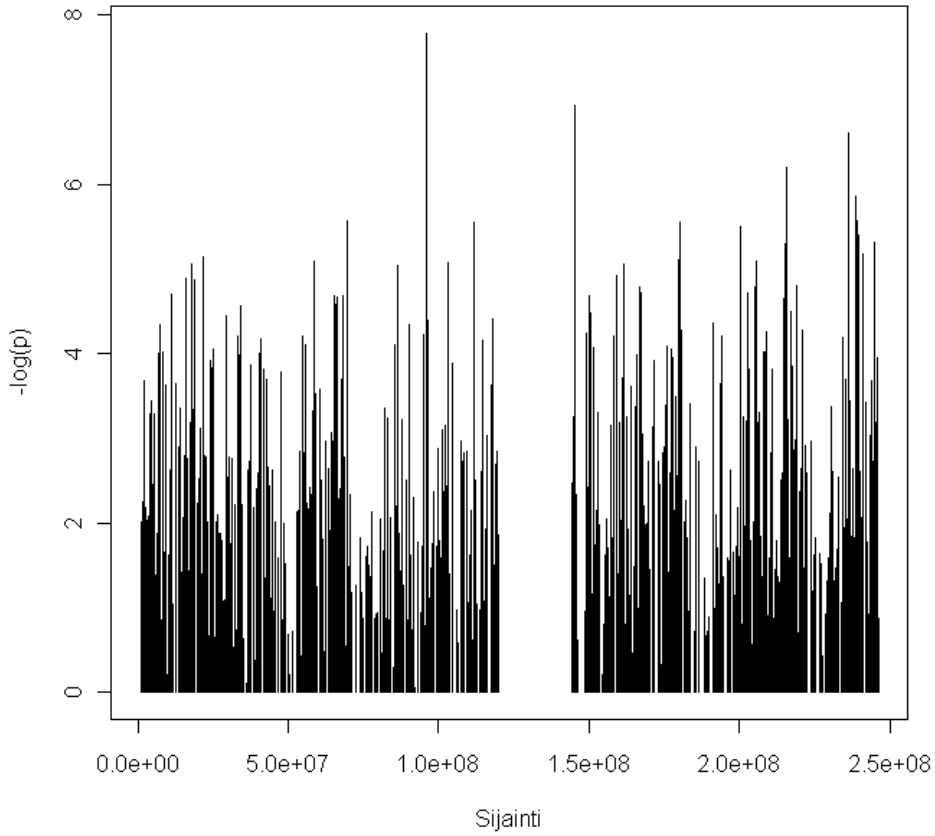
Blokeille tehtiin logistinen regressioanalyysi ja tuloksena saatiin taulukon 9 mukainen tuloslista. Blokit on järjestetty kromosomin numeron mukaan. BP1-sarake kertoo, mistä kohtaa

blokki alkaa ja BP2 mihin se loppuu. Testisuureena on Waldin testisuure. Myös näiden tulostaulukoiden perusteella piirrettiin p-arvokuvaajat, kuten kuva 7.

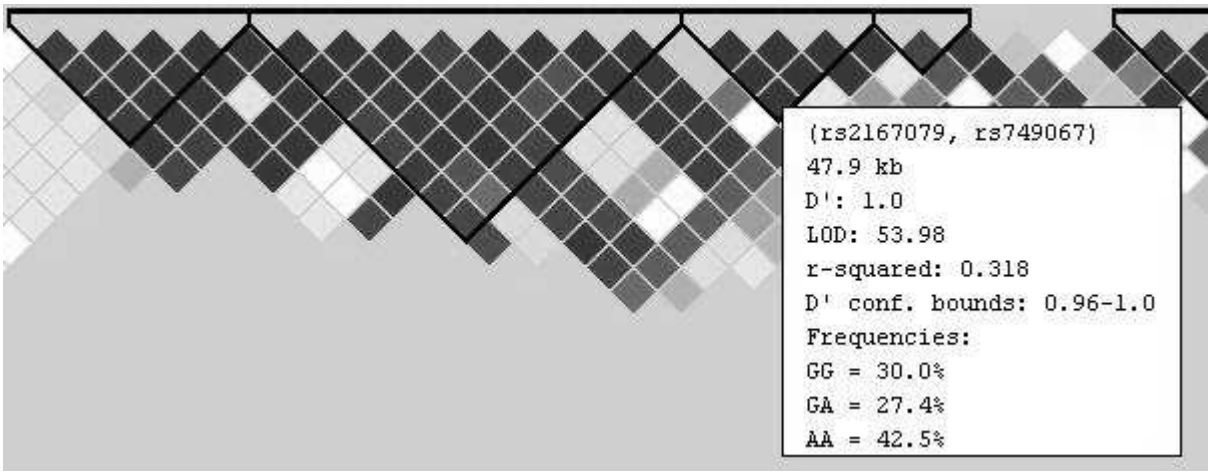
Taulukko 9. Haplotyyppien logistisen regressioanalyysin tuloksia.

Kromosomi	BP1	BP2	Haplotyyppi	Testisuure	p-arvo
1	2310562	2328934	AAA	5,01	0,0252
1	4160278	4169351	GC	4,34	0,0373
1	4452662	4458793	GAG	3,9	0,0483
1	4452662	4458793	GGG	4,6	0,032
1	5427140	5436781	AG	4,33	0,0375
1	5675937	5687940	GAA	4,2	0,0404
1	6963762	6977611	GGAA	3,85	0,0498
1	6963762	6977611	AGGG	5,59	0,0181
1	7102308	7128537	GGGAG	4,43	0,0353
1	7238701	7240128	GG	6,19	0,0129

Koska logistinen regressioanalyysi testaa kaikkien blokkien eri haplotyypit, on tuloslistassa yhteensä 54 451 haplotyyppiä, joista osa on saman blokin eri haplotyyppejä. Näistä merkitseviä merkitsevyysasteella 0,05 on 3 060 haplotyyppiä. Blokkeja analysoitiin yhteensä 16 277 ja näistä merkitseviä merkitsevyysasteella 0,05 oli 2 445 ja merkitsevyysasteella 0,01 merkitseviä oli 564. Suhteellisesti eniten merkitseviä blokkeja on kromosomeissa 11 (p-arvokuvaaja liitteessä 3), 22 ja 23. Kuvassa 8 on neljä peräkkäistä voimakkaan LD:n blokkia kromosomista 11 ja kolmannen blokin kärkiparin, SNP-markkereiden rs2167079 ja rs749067, LD:n mittasuureet sekä haplotyyppien frekvenssit.

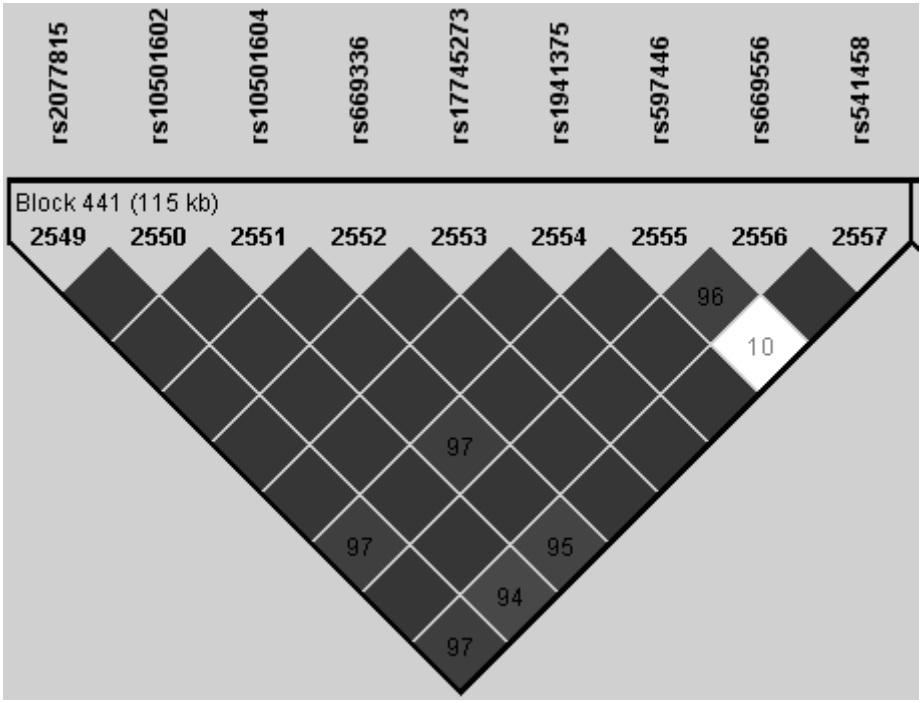


Kuva 7. Kromosomin 1 blokkien p-arvokuvaaja.



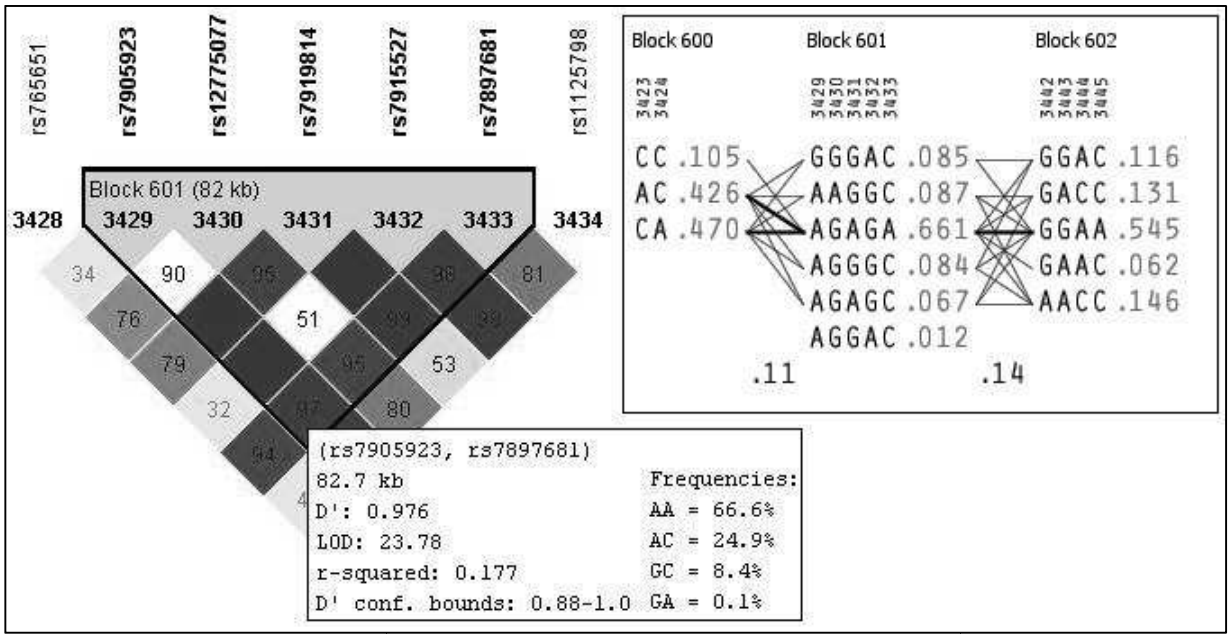
Kuva 8. Kromosomissa 11 voimakkaan LD:n blokkeja. Laatikossa kolmannen lohkon kärkiparin testisuureita.

Kromosomista 11 löytyi yhteensä 175 merkittävää haplotyyppiä merkitsevyysasteella 0,05. Kuvassa 9 on yksi suurimmista blokeista. Siinä on 9 SNP-markkeria ja 8 eri haplotyyppiä. Haplotyypeistä kaksi on merkitseviä 0,05 merkitsevyysasteella, AAAGAAGAA ($p = 0{,}00861$) ja AAGGGAAAA ($p = 0{,}0391$). Voidaan siis olettaa, että blokilla on vaikutusta syövän syntymiseen, ja blokki ja sen ympäristö kromosomissa olisi sopiva jatkotutkimuksen kohde.



Kuva 9. Kromosomin 11 haplotyyppiblokki.

Kun kaikki kromosomit otetaan huomioon ja samojen blokkien eri haplotyypeistä oli poistettu muut paitsi merkitsevin haplotyyppi, merkitsevyysasteen 0,001 alapuolelle jäi 61 haplotyyppiä. Näistä haplotyypeistä 15 merkitsevintä on taulukossa 10. Siinä on ilmoitettu SNP-markkereiden lukumäärä blokissa (N), blokin kromosomi, blokin alkamis- ja päättymislokukset, blokin ensimmäinen SNP, haplotyyppi, Waldin testisuure haplotyypille sekä testin p-arvo. Kuvassa 10 on merkitsevimmän blokin (601) kuva *Haploview*'llä piirrettynä.



Kuva 10. Haploview-kuva merkitsevimmästä haplotyyppiblokista logistisen regressioanalyysin perusteella.

Kuvassa 10 on haplotyyppiblokin kuvan lisäksi tiedot blokin eri haplotyypeistä sekä vierekkäisistä blokeista. Blokin kärkiparista (rs7905923, rs7897681) on kytkentäepätasapai-

nomääreet ja alleelien frekvenssit merkittynä alimpaan laatikkoon. Oikeanpuoleisessa laatikossa on blokin haplotyypit frekvensseineen sekä kytkennät vierekkäisiin blokkeihin ja niiden haplotyyppeihin. SNP-markkerit on merkitty numeroin. Viereisissä blokeissa ei ollut merkitseviä haplotyyppejä logistisen regressioanalyysin tulosten perusteella, mutta blokin 601 alleelin AGAGA p-arvo on alle 0,05 (0,011).

Taulukko 10. Logistisen regressioanalyysin mukaan 15 merkitsevintä haplotyyppiblokkia.

N	Kromo	Alkulokus	Loppulokus	SNP1	Haplotyyppi	Testisuure	p-arvo
5	10	109663117	109745886	rs7905923	GGGAC	20,36	6,20E-07
3	9	74485413	74509273	rs1417614	AAG	19,31	9,00E-07
2	21	35786327	35803265	rs7281771	AA	15,58	1,60E-06
4	9	80632505	80670516	rs11137862	AAAA	16,45	1,80E-06
2	22	33744017	33757427	rs5750033	CG	15,39	2,40E-06
2	2	37889932	37890802	rs10865127	AA	16,74	4,20E-06
2	2	142125079	142131001	rs1437352	GG	15,29	4,80E-06
3	17	28589889	28616992	rs9972931	GAA	16,93	8,70E-06
2	2	233695966	233707325	rs10193128	GA	15	0,000105
3	11	63881718	63906946	rs541207	AGG	15	0,00011
2	2	114149360	114166498	rs7575011	AG	14,7	0,000129
2	21	32716387	32717832	rs2833801	GA	14,6	0,000136
2	2	142125079	142131001	rs1437352	AG	14,5	0,000137
3	2	205739367	205743311	rs17188399	GAG	14,5	0,000138
2	5	41267570	41270400	rs4957381	AG	14,4	0,000147

Taulukkoa 10 tarkasteltaessa on syytä huomata, että 15 merkitsevimmän haplotyypin joukossa on peräti kuusi haplotyyppiä kromosomista 2. Kaksi haplotyypeistä kuuluu samaan blokkiin (GG ja AG aloitusmarkkerilla rs1437352), joten tällä blokilla voidaan olettaa olevan vaikutusta syövän syntymiseen. Blokkien lähekkäisyyteen on myös syytä kiinnittää huomiota, sillä alueella voi olla syövän kehittymisen kannalta tärkeä geeni. Kromosomin 2 haplotyypeistä lähekkäin ovat blokit aloittavilla SNP-markkereilla rs7575011 ja rs1437352 sekä rs17188399 ja rs10193128. Myös kromosomeilla 9 ja 21 on useampi haplotyyppi 15 merkitsevimmän joukossa ja molempien haplotyyppiblokit ovat kohtalaisen lähekkäin.

4.3 Random Jungle -tuloksia

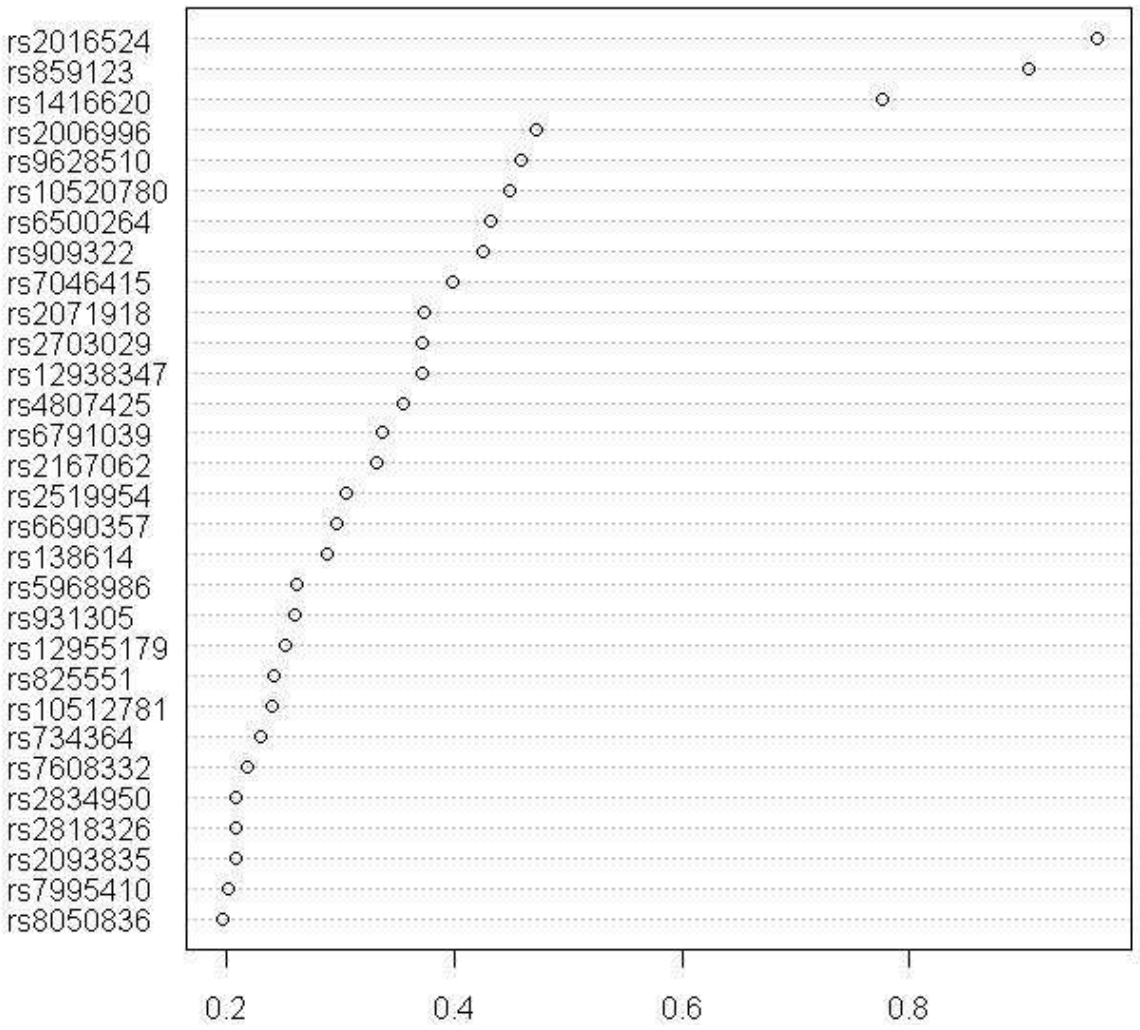
Data jaettiin aluksi kolmeen osaan laskenta-aikojen lyhentämiseksi. Jokaisessa osassa on noin 30 000 SNP-markkeria. Lisäksi generoitiin yhdeksän eri opetus- ja testausaineistoa, jotka ovat ns. tasapainotettuja, eli kontrollitapauksia on yhtä paljon kuin sairastuneita. Sairastuneiden ja verrokkitapausten suuri lukumäärällinen ero aiheutti sen, ettei *Random Jungle* tehnyt onnistuneita luokitteluja ilman tasapainotusta, vaan ennusti kaikki tapaukset terveiksi. Koska RJ ei pysty käsittelemään tyhjiä arvoja, jätettiin sellaisia sisältävät muuttujat opetusaineistoista pois. Ensimmäisen kierroksen jälkeen valittiin mukaan ne SNP-markkerit, joiden tärkeysindeksi on suurempi kuin 0. Aineistot yhdistettiin ja markkereiden valintakriteeriä, eli tärkeysmitan suuruutta, korotettiin joka kierroksen jälkeen.

Taulukko 11. Esimerkki RF-puusta.

Vasen oksa	Oikea oksa	Jakomuuttuja	Jakopiste	Päättävä solmu	Ennuste
2	3	rs541207	3	1	-
4	5	rs6690357	3	1	-
6	7	rs2016524	7	1	-
8	9	rs4807425	1	1	-
10	11	rs10263926	3	1	-
12	13	rs10520780	1	1	-
14	15	rs825551	4	1	-
0	0	-	0	0	1

Taulukossa 11 on esimerkki yhdestä RF-algoritmin muodostaman puun alkuosasta. Vasen ja oikea oksa kertovat, millä rivillä seuraavat solmukohdat ovat kyseisen solmun markkeriin verrattuna. Jakomuuttuja on solmukohtaan valittu markkeri. Jakopiste kertoo, miten SNP-markkerin alleelit on jaoteltu. Alleelit on muunnettu binäärimuuttujiksi, joten jos jakopiste on 4, binäärimuuttuja on muotoa (0, 0, 1), sillä $0*2^0 + 0*2^1 + 1*2^2 = 4$. Se siis tarkoittaa, että kahden ensimmäisen alleelin havainnot ovat jakautuneet vasemmalle ja viimeinen oikealle. Vasemmalle jakautuvat havainnot kuuluvat todennäköisemmin sairastuneiden luokkaan kuin oikealle jakautuvat. Päättävä solmu -sarake kertoo, onko kyseinen solmu päättävä siten, että mikäli arvo on 1, solmu ei ole päättävä. Ennuste-sarakkeessa on päättävälle solmulle laskettu ennuste siitä luokasta, mihin se todennäköisesti kuuluu. Viimeisen rivin solmu on siis päättävä ja sille laskettu luokkaennuste on 1 eli sairastuneiden luokka.

Kun SNP-markkereita oli jäljellä noin 3 000, generoitiin yhdeksän opetus- ja testausaineistoa lisää. Lopulta päädyttiin 72 SNP-markkeriin, joille tehtiin myös χ^2-testi. Kuvassa 11 on valittujen SNP-markkereiden tärkeysindeksikuvaaja, jossa vaaka-akselilla on prosenttiluku mikä markkerin saama tärkeysindeksin arvo on maksimiin verrattuna. Taulukossa 12 on 15 pienimmän p-arvon saanutta SNP-markkeria sekä niiden sijainti, tärkeysindeksi sekä esiintymä. Esiintymä on niiden tuloslistojen lukumäärä, jossa kyseinen markkeri oli tärkeimpien joukossa, kun kaikki 18 opetusaineistoa oli analysoitu. Mitä suurempi markkerin esiintymä on, sitä useammalla testausaineistolla se valikoitui tärkeimpien joukkoon. Seitsemällä SNP-markkerilla χ^2-testin p-arvo oli suurempi kuin 0,01, mutta kaikki olivat alle 0,02.



Kuva 11. RF-algoritmin valitsemien muuttujien tärkeysindeksikuvaaja.

Taulukko 12. RF-algoritmin valitsemien SNP-markkereiden 15 merkitsevintä χ^2-testin perusteella.

SNP	Kromosomi	BP	Esiintymä	Tärkeysindeksi	P-arvo
rs2016524	18	63488431	6	2,26313	3,03E-09
rs859123	12	25101031	5	5,57218	4,53E-09
rs7995410	13	47651319	15	5,45629	7,23E-06
rs4807425	19	3180224	10	3,45385	1,31E-05
rs319749	17	28596020	7	3,35128	1,97E-05
rs2519954	7	88240421	8	4,68837	3,41E-05
rs12938347	17	65451592	10	3,78746	3,49E-05
rs7915527	10	109721574	5	3,02258	4,14E-05
rs1416620	1	214963754	8	6,07325	4,25E-05
rs7046415	9	80670516	14	3,0005	5,01E-05
rs2006996	9	116632459	5	3,36715	5,25E-05
rs11926147	3	54170136	8	2,5906	7,46E-05
rs9628510	22	47425790	4	2,69793	0,000127
rs2834950	21	35803265	6	3,09139	0,000236
rs931305	4	186711694	7	8,10189	0,000241

72 SNP-markkerin joukossa oli SNP-markkereita lähes jokaisesta kromosomista, eniten kromosomista 2 (7 kpl) ja vähiten kromosomeista 14 ja 20 (0 kpl). Taulukossa 13 on tarkemmin eritelty SNP-markkereiden jakautuminen kromosomeihin. Taulukkoon on laskettu myös joka kromosomista valittujen SNP-markkereiden prosenttiosuus kaikista valituista. Kromosomit 14 ja 20 jätettiin taulukosta pois, sillä niistä ei valittu yhtään SNP-markkeria. Keskimäärin joka kromosomista *Random forest* -algoritmi valitsi 4 SNP-markkeria.

Taulukko 13. Random forest -algoritmin valitsemien SNP-markkereiden jakautuneisuus kromosomeihin.

Kromo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**15**	**16**	**17**	**18**	**19**	**21**	**22**	**23**
lkm	4	7	5	3	3	5	4	2	6	2	4	2	2	2	3	2	5	1	2	3	5
%	5,6	9,7	6,9	4,2	4,2	6,9	5,6	2,8	8,3	2,8	5,6	2,8	2,8	2,8	4,2	2,8	6,9	1,4	2,8	4,2	6,9

Lähekkäiset SNP-markkerit tarkistettiin sen varalta, kuuluvatko ne samaan blokkiin, ja valittujen SNP-markkereiden joukosta löytyi kolme *plink*-ohjelmalla muodostettua blokkia, joista jokaisesta oli kaksi SNP-markkeria algoritmin tuloslistassa. Kahdella blokilla oli ainakin yksi haplotyyppi, jonka logistisen regressioanalyysin perusteella laskettu p-arvo oli pienempi kuin 0,01. Blokkien merkitsevimmät haplotyypit ovat taulukossa 14.

Taulukko 14. Random Jungle -algoritmin avulla löydettyjen SNP-markkereiden joukossa olevat blokit.

SNP lkm	**Kromosomi**	**Alkulokus**	**Loppulokus**	**Ensimmäinen SNP**	**Viimeinen SNP**	**Haplotyyppi**	**P-arvo**	**Mukana olevat SNP:t**	
6	1	214945081	214979884	rs1416611	rs2813703	AGGCGG	0,006	rs1416620	rs2995381
3	13	47143354	47152997	rs7990134	rs9567906	GGA	0,161	rs7995410	rs9534995
2	21	35786327	35803265	rs7281771	rs2834950	AA	2E-06	rs7281771	rs2834950

Tärkeimmiksi havaittujen SNP-markkereiden hajanaisuus ympäri genomia oli positiivinen yllätys. Algoritmi onnistui poimimaan hyvin sellaisia SNP-markkereita, jotka eivät olleet suuressa riippuvuudessa keskenään. Tämä on merkittävä etu yksittäisten SNP-markkereiden vaikutusta testaaviin menetelmiin, kuten χ^2-testiin, jotka eivät huomioi SNP-markkereiden keskinäisiä riippuvuuksia. Menetelmän ongelmana on kuitenkin herkkyys aineiston epätasaisuudelle ja tyhjien arvojen huono sietokyky, sillä näiden ongelmien korjaamiseksi jouduttiin analyysin ulkopuolelle jättämään sekä SNP-markkereita että kontrollitapauksia, ja on mahdollista, että jokin merkittävä markkeri on jäänyt algoritmia varten valitun markkeriotoksen ulkopuolelle.

Yllä kuvattu ongelma voitaisiin tulevaisuudessa välttää huolellisella aineistonkeruulla, jossa otettaisiin algoritmin vaatimukset tapaus- ja verrokkiryhmien koon yhtäläisyydestä huomioon. Tämä asettaa aineiston keruulle monenlaisia haasteita, jotta aineistoa koostettaessa ei tulisi tehtyä valintaa, joka saa aikaan harhaa. Myös tyhjiä arvoja tulisi mahdollisuuksien mukaan välttää. Vaikka esimerkiksi Breimanin et al. 2010 kehittämässä *R*-ohjelman *randomForest* -paketissa on menetelmä tyhjien arvojen täyttämiseksi mm. havainnon keskiarvolla tai mediaanilla, olisi tulosten luotettavuuden kannalta parempi välttää tällaisia menetelmiä ja mieluummin jättää tyhjiä arvoja sisältävät muuttujat analyysin ulkopuolelle.

5 Johtopäätökset

SNP-markkereiden valtavan määrän takia oli odotettavissa, että kaikki menetelmät löytäisivät merkitseviä markkereita. Tuloksia lukiessa onkin oltava kriittinen, sillä ns. vääriä positiivisia on joukossa paljon. Hylkäysvirheen kontrollointimenetelmiä on tarjolla perinteisille analyysimenetelemille ja sellaista käytettiinkin χ^2-testin ja logistisen regressioanalyysin p-arvojen korjaamiseksi. Voidaan kuitenkin olettaa, että ne markkerit, jotka olivat kaikkien analyysimenetelmien tuloslistojen kärkipäässä, ovatkin todennäköisimmin oikeasti merkitseviä. Jatkotutkimuksia ajatellen SNP-markkereiden joukkoa on onnistuttu supistamaan.

Sekä χ^2-testin että logistisen regressioanalyysin perusteella yli tuhat SNP-markkeria olivat erittäin merkitseviä. Nämä testit eivät kuitenkaan ota huomioon SNP-markkereiden keskinäisiä riippuvuussuhteita eivätkä yhteisvaikutusta sairauden puhkeamiseen. Tuhat SNP-markkeria on lisäksi todella suuri joukko lähteä analysoimaan tarkemmin. Tämänkaltaiset testit eivät siis yksinään riitä erottelemaan niitä tekijöitä, joilla on oikeasti merkitystä sairastuvuuteen. Ne kuitenkin kaventavat tutkittavaa joukkoa, ja varsinkin p-arvokuvaajien avulla voidaan silmämääräisesti arvioida, mistä kohtaa genomia mahdolliset vaikuttavat geenit tai säätelyalueet sijaitsevat. Hylkäysvirhekontrollin jälkeen huomattiin, että oikeasti merkitseviä SNP-markkereita olikin vain muutama. Tämä osoittaa, kuinka yksinkertaisen assosiaatiotestit eivät ole yksistään sopivia riippuvaisten SNP-markkereiden analyysiä varten.

Haplotyyppiblokkien muodostaminen poistaa SNP-markkereiden keskinäisen riippuvuuden ongelman. Kun hyvin vahvassa riippuvuussuhteessa olevat lähekkäiset SNP:t on saatu niputettua yhteen, on tutkittava joukko jälleen pienentynyt. Kuitenkin blokkien testauksen jälkeen merkitseviä blokkeja oli yli 2 000. Kaikkien blokkien ja niiden eri haplotyyppien läpikäyminen olisi edelleen valtavan työlästä. *Haploview*'n avulla piirretyt kuvat antavat suunnan siitä, miten blokit ovat kromosomissa jakautuneet, mutta eivät kerro missä kohtaa ovat merkitsevimmät haplotyypit. P-arvokuvaajat ovat käyttökelpoisia näidenkin tulosten silmämääräisessä tarkastelussa ja verrattuna yksittäisten assosiaatiotestien p-arvokuvaajiin, merkitsevät alueet ovat huomattavasti vähentyneet.

Tutkittavaa SNP-joukkoa on siis kytkentäepätasapainomenetelmällä jälleen saatu pienemmäksi, mutta tarkempia jatkotutkimuksia ajatellen SNP-markkereita on silti liikaa. Vaikka jokaisesta blokista valitsisi vain yhden markkerin, on tutkittava joukko silti valtava. Blokkien muodostaminen on kuitenkin yksi sopivimmista analyysimenetelmistä laajojen SNP-aineistojen tutkimiseen, ja yhdistettynä joihinkin muihin menetelmiin, kuten perinteiseen χ^2-testiin, voidaan saada luotettavia tuloksia.

Random forest -algoritmi onnistui hyvin poimimaan aineistosta markkereita, jotka eivät ole riippuvaisia toisistaan. 72 tärkeimmän SNP-markkerin joukossa oli vain kolme paria, jotka olivat kytkentäepätasapainomääritelmän mukaan riippuvaisia toisistaan. Tulos on hyvin rohkaiseva, sillä algoritmi saa tiivistettyä laajankin aineiston tehokkaasti. Ongelmana on aineiston tasapainotus, sillä algoritmi ei toimi, mikäli verrokkitapauksia on moninkertainen määrä sairastuneisiin nähden. Tämä oli tilanne myös tutkielman aineiston kanssa, ja suuri joukko kontrollitapauksia jouduttiin jättämään jokaisen opetus- ja testausaineiston ulkopuolelle. Tämä voitaisiin huomioida tulevaisuudessa jo aineiston keruun aikana. Tässä tutkielmassa ongelmaa yritettiin pienentää generoimalla useita eri opetus- ja testausaineistoja (18 kpl), joten jokaisella kontrollitapauksella oli kohtalaiset mahdollisuudet päästä analyysiin

mukaan. Voidaan siis olettaa, että valitut SNP-markkerit ovat merkitseviä ja vertailut perinteisiin menetelmiin tukevat tätä ajatusta.

Verrattaessa eri menetelmien tuloslistojen 15 merkitsevintä SNP-markkeria (taulukot 5, 10 ja 12 sekä liite 1), huomataan jonkun verran yhtäläisyyksiä. Haplotyyppiblokkien kärkijoukosta valittiin jokaisen blokin ensimmäinen SNP-markkeri. Kahdelle eri tuloslistalle ovat päässeet 9 markkeria ja lähes kaikki ovat sekä logistisen regressioanalyysin että χ^2-testin tuloslistojen kärkipäässä. Kaksi SNP-markkeria on kolmella tuloslistalla (rs10193128, kromosomi 2 ja rs7905923, kromosomi 10) ja yksi markkeri (rs541207, kromosomi 11) on kaikkien menetelmien mukaan merkitsevä. Valitsemalla suuremman joukon tuloslistojen merkitsevimpiä markkereita yhtäläisyyksiä tulee vastaavasti enemmän. Näiden SNP-markkereiden joukosta 12 on sekä χ^2-testin että logistisen regressioanalyysin 15 merkitsevimmän SNP-markkerin joukossa ja tuloslistat olivat muutenkin hyvin samanlaisia, joten molempien testien käyttäminen yhtä aikaa ei ole järkevää.

Taulukossa 15 on 15 SNP-markkeria, jotka ovat olleet vähintään kolmen eri menetelmän tuloslistojen 40 merkitsevimmän joukossa. Taulukossa on ilmoitettu SNP-markkerin kromosomi, mahdollisen haplotyyppiblokin p-arvo, χ^2-testin p-arvo, logistisen regressioanalyysin p-arvo sekä *Random forest* -algoritmin tärkeysindeksin arvo. Listasta on poistettu ne SNP-markkerit, jotka eivät ole RF-algoritmin valitsemia (7 kpl). Mikäli SNP-markkerille on merkitty haplotyyppiblokin p-arvo, se on joko blokin ensimmäinen tai viimeinen SNP, eli ne joilta haplotyyppiblokin tulos puuttuu, voivat silti olla osana haplotyyppiblokkia. Assosiaatiotestien p-arvoja ei yhteenvetotaulukoihin ole korjattu *FDR*-menetelmällä.

Taulukko 15 Eri menetelmien tulosten yhtäläisyyksiä.

Kromo.	SNP	LD-blokit	χ^2-testi	Log.reg.	Random forest
2	rs10193128	0,000105	4,68E-05	6,20E-05	3,11726
2	rs1437352	4,80E-06	0,0001209	0,0001487	3,86258
3	rs6791039	-	4,83E-05	0,0001167	3,68753
4	rs931305	-	0,0001573	0,0002598	4,29957
5	rs10512781	0,000147	9,92E-05	0,0001472	3,53857
6	rs1744173	-	0,0001446	0,0002369	3,09684
8	rs734364	-	0,000131	0,0001658	3,34946
9	rs2006996	-	0,0001267	0,0003013	4,50474
9	rs7046415	1,80E-06	2,31E-05	5,18E-05	4,58793
10	rs7915527	-	7,58E-06	2,27E-05	4,74735
11	rs541207	0,00011	3,78E-05	4,34E-05	3,45385
13	rs7995410	-	7,87E-05	0,0001707	6,78247
17	rs319749	-	5,44E-06	2,00E-05	5,57218
19	rs4807425	-	1,96E-06	5,74E-06	6,07325
21	rs2834950	1,60E-06	3,11E-05	8,16E-05	4,31117

Tulosten yhtäläisyydet ovat melko rohkaisevia. Taulukon 15 SNP-markkerit ovat kaikki lisätutkimuksen arvoisia ja niiden voidaan olettaa vaikuttavan sairauden puhkeamiseen. Menetelmien erilaisuudesta huolimatta *Random forest* -algoritmin mukaan tärkeimmät SNP-markkerit ovat myös merkitseviä perinteisten menetelmien mukaan. Se on osoitus algoritmin toimivuu-

desta. Voidaan siis todeta, että sekä RF-algoritmi että myös haplotyyppiblokkien muodostaminen ovat sopivia menetelmiä laajan SNP-aineiston tutkimiseen.

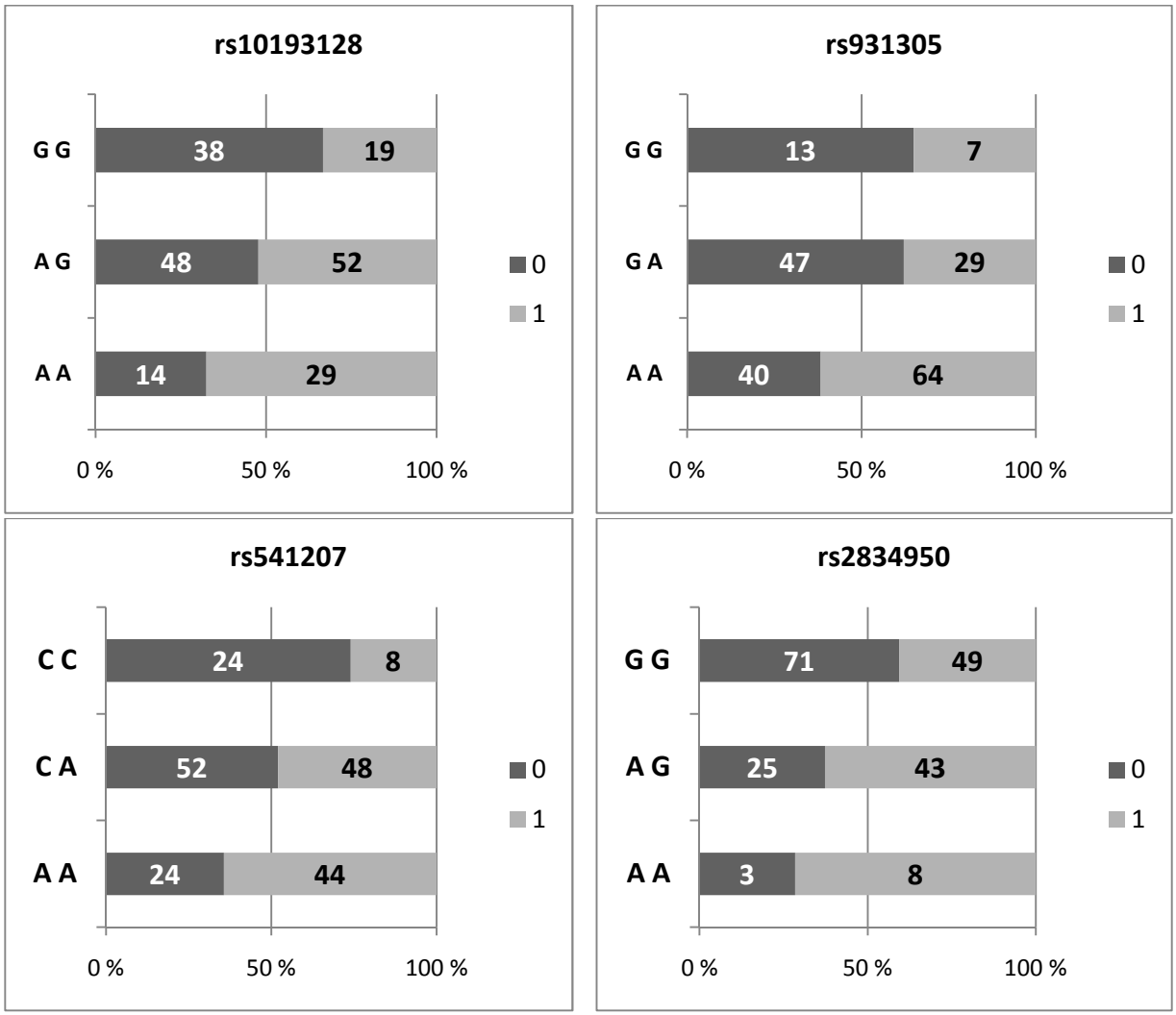
Taulukossa 16 on kromosomikohtainen yhteenveto tuloksista muiden menetelmien kuin *Random forest* -algoritmin osalta. RF -algoritmin kromosomikohtaiset tulokset ovat luvussa 4.3 taulukossa 13. Taulukossa 16 on laskettu kromosomin tilastollisesti merkitsevien SNP-markkereiden lukumäärä logistisen regressioanalyysin perusteella merkitsevyystason ollessa 0,05 ilman *FDR*-kontrollia ja merkitsevien SNP-markkereiden osuus kromosomin kaikkien markkereiden lukumäärästä. Kromosomikohtaiset markkereiden kokonaislukumäärät voivat tarkistaa kuvasta 3 luvusta 2.3. Lisäksi on laskettu kromosomiin muodostettujen haplotyyppiblokkien lukumäärä, haplotyyppien lukumäärä, merkitsevien haplotyyppien lukumäärä logistisen regressioanalyysin perusteella sekä merkitsevien haplotyyppien prosentuaalinen osuus kaikkien haplotyyppien lukumäärästä.

Taulukko 16 Yhteenveto kromosomikohtaisista tuloksista.

Kromo	Merkitsevät SNPt	Merkitsevien %osuus	Blokkeja	SNPiden %osuus blokeissa	Haplot. lkm	Merkitsevät haplot. lkm	Merkitsevien %osuus
1	386	5,651	1222	17,89	4076	207	5,079
2	379	5,121	1285	17,36	4339	224	5,162
3	335	5,892	968	17,02	3252	148	4,551
4	312	6,621	773	16,40	2547	157	6,164
5	353	6,163	1054	18,40	3531	215	6,089
6	379	6,510	1093	18,77	3719	228	6,131
7	276	5,628	839	17,11	2833	149	5,259
8	317	5,597	1097	19,37	3754	194	5,168
9	274	6,420	780	18,28	2601	150	5,767
10	260	5,897	797	18,08	2710	148	5,461
11	297	6,901	762	17,70	2530	171	6,759
12	278	6,506	773	18,09	2577	154	5,976
13	154	5,135	519	17,31	1733	94	5,424
14	164	5,712	488	17,00	1610	87	5,404
15	161	5,681	510	18,00	1699	78	4,591
16	166	6,173	463	17,22	1511	97	6,420
17	157	5,836	462	17,17	1555	84	5,402
18	164	6,072	477	23,03	1581	74	4,681
19	119	6,471	282	15,33	935	57	6,096
20	150	6,242	429	17,85	1401	81	5,782
21	112	6,854	336	20,56	1104	69	6,250
22	100	6,281	295	18,53	998	67	6,713
23	221	6,475	572	16,76	1855	127	6,846
Yht.	5514	6,057	16276	17,88	54451	3060	5,620

Taulukosta 16 voi havaita, että kromosomissa 11 on selvästi keskiarvoa enemmän merkitseviä SNP-markkereita ja haplotyyppejä. Muita keskimääräistä enemmän merkitseviä tuloksia saaneita kromosomeja ovat 4, 21 ja 23. Keskimääräistä vähemmän merkitseviä tuloksia on kro-

mosomeissa 1 ja 2, joissa on kuitenkin selvästi muita enemmän SNP-markkereita. Myös kromosomit 7 ja 8 saivat keskimääräistä vähemmän merkitseviä tuloksia siitä huolimatta, että niiden SNP-markkereiden lukumäärä on hyvin lähellä keskiarvoa (n. 4 000 SNP-markkeria). Huomattavaa on, että *Random forest* -algoritmi valitsi eniten SNP-markkereita kromosomista 2 ja keskimääräisesti kromosomista 11. Kromosomikohtaisia tuloksia tarkastelemalla menetelmät eivät siis anna yhdenmukaisia tuloksia.



Kuva 12. Merkitsevimpiä SNP-markkereita ja niiden terveiden ja sairastuneiden frekvenssit alleeleittain.

Kuvassa 12 on neljä taulukosta 15 poimittua SNP-markkeria graafisesti esitettynä. Kuvasta näkyy, kuinka terveiden ja sairastuneiden osuudet vaihtelevat eri alleelien väleillä. Vaakapalkkien keskellä olevat numerot ovat terveiden ja sairastuneiden havaintojen frekvenssit. Tummempi väri tarkoittaa fenotyyppiä 0 eli tervettä havaintoa, ja vaaleampi sairastunutta. Huomattavaa on erityisesti SNP-markkerilla rs931305 se, että yleisimmällä alleelilla (AA) sairastuneiden osuus on huomattavasti suurempi kuin terveiden.

Tulosten perusteella kromosomia 11 olisi syytä tutkia tarkemmin ja erityisesti SNP-markkerin rs541207 ympäristöstä. Myös kromosomin 21 markkerin rs2834950 ympäristössä voisi olla syövän syntymiseen vaikuttavia tekijöitä. Näiden lisäksi mahdollisia syöpäalttiuteen vaikuttavia alueita voi olla merkitsevimpien haplotyyppien alueella (taulukko 8), näissä erityi-

sesti kromosomien 2, 9 ja 21 lähekkäisten haplotyyppien vaikutus olisi hyvä selvittää. Myös yhteenvetotaulukosta 12 olisi kannattavaa tutkia ainakin sellaiset SNP-markkerit, jotka ovat myös osana haplotyyppiblokkia.

Tässä tutkielmassa käytetyt analyysimenetelmät onnistuivat vaihtelevasti valikoimaan laajasta joukosta SNP-markkereita sellaisia, jotka vaikuttaisivat syövän syntymiseen. Huomattiin, että erityisesti sellaiset perinteiset menetelmät, joissa oletuksena on muuttujien riippumattomuus, eivät olleet sopivia SNP-markkereiden tutkimiseen. Tällaisia menetelmiä voidaan kuitenkin käyttää uudempien menetelmien rinnalla. Myös uudemmissa menetelmissä oli heikkouksia, vaikka ne ovatkin suunniteltu SNP-aineistojen tutkimista varten. Kuitenkin haplotyyppiblokkien muodostaminen ja *Random forest* -algoritmin kaltaiset menetelmät ovat lupaavia SNP-aineistojen analysointia varten. Uusia menetelmiä kuitenkin julkaistaan jatkuvasti, joten kriittisiä tarkasteluja olisi syytä tehdä tulevaisuudessakin.

Lähteet

Agresti, A. (1990): ”Categorical Data Analysis”, John Wiley and Sons, Inc., USA.

Akey, J., Jin, L. & Xiong, M. (2001), ”Haplotypes vs single marker linkage disequilibrium tests: what do we gain?”, *European Journal of Human Genetics*, 9, 291–300, Nature Publishing Group, Iso-Britannia.

Barret, J. C., Fry, B., Maller, J. & Daly, M. J. (2005), ”Haploview: analysis and visualization of LD and haplotype maps”, *Bioinformatics*, 21:2, 263–265, Oxford University Press, Iso-Britannia. Saatavilla Internetistä: http://www.broad.mitt.edu/mpg/haploview

Benjamini, Y. & Hochberg, Y. (1995), ”Controlling the false discovery rate: A practical and powerful approach to multiplle testing”, *Journal of the Royal Statistical Society*, Series B (Methodological), 57:1, 289–300, Iso-Britannia.

Botstein, D. & Risch, N. (2003), ”Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex diseases”, *Nature Genetics*, 33, 228–237, Nature Publishing Group, Iso-Britannia.

Breiman, L. (1996), ”Bagging predictors”, *Machine Learning*, 24, 123–140, Kluwer Academic Publishers, Boston.

Breiman, L. (2001), ”Random Forest”, *Machine Learning*, 45, 5–32, Kluwer Academic Publishers, Boston.

Breiman, L., Cutler, A., Liaw, A. & Wiener, M. (2010), ”Package ’randomForest”. Saatavilla Internetistä: http://cran.r-project.org/web/packages/randomForest/index.html

Cantor, R. M., Lange, K. & Sinsheimer, J. S. (2010), ”Prioritizing GWAS results: A review of statistical methods and recommendations for their application”, *The American Journal of Human Genetics*, 8, 6–22, Elsevier Inc., USA.

Collins, F.S., Morgan, M. & Patrinos, A. (2003), ”Human Genome Project: Lessons from large-scale biology”, *Science* (Vol. 300, no. 5617), 286–290, American Association for the Advancement of Sience, USA.

Deng, H.-W., Li, J. & Recker, R. R. (2001), ”LOD score exclusion analyses for candidate genes using random population samples”, *Annals of Human Genetic*, 65:3, 313-329, Iso-Britannia.

Efron, B. (1979), ”Bootstrap methods: Another look at the Jackknife”, *The Annals of Statistics* (Vol. 7, no. 1), 1–26, Institute of Mathematical Statistics, USA.

Engle, L. J., Simpson, C. L. & Landers, J. E. (2006), ”Using high-throughput SNP technologies to study cancer”, *Oncogene*, 25, 1594–1601, Nature Publishing Group, Iso-Britannia.

Gabriel, S. B., Scaffner, S.F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. & Altshuler, D. (2002), ”The Structure of Haplotype Blocks in the Human Genome”, *Science* (Vol. 296, no. 5576), 2225–2229, American Association for the Advancement of Sience, USA.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), ”The Elements of Statistical Learning: Data Mining, Inference and Prediction”, Springer-Verlag, New York.

Lewontin, R. C. & Kojima, K. (1960), "The evolutionary dynamics of complex polymorphisms", *Evolution*, 14, 450–472, Society for the Study of Evolution.

Lunetta, K. L., Hayward, L. B., Segal, J. & van Eerdewegh, P. (2004), ”Screening large-scale association study data: exploiting interactions using random forests”, *BMC Genetics*, 5:32, BioMed Central Ltd.

Ollikainen, V. & Uimari, P. (2004), ”Geenikartoitusopas”, *CSC – Tieteellinen laskenta Oy*, Espoo.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J. & Sham, P. C. (2007), ”PLINK: a toolset for whole-genome association and population-based linkage analysis”, *American Journal of Human Genetics*, 81, 559–575, Elsevier Inc., USA. Saatavilla Internetistä: http://pngu.mgh.harvard.edu/purcell/plink/

Schwarz, D. F., König, I. R. & Ziegler, A. (2010), ”On Safari to Random Jungle: A fast implementation of Random Forest for high dimensional data”, *Bioinformatics* (Vol. 26, 14th ed.), 1752–1758, Oxford University Press. Saatavilla Internetistä: http://www.randomjungle.org/rjungle/

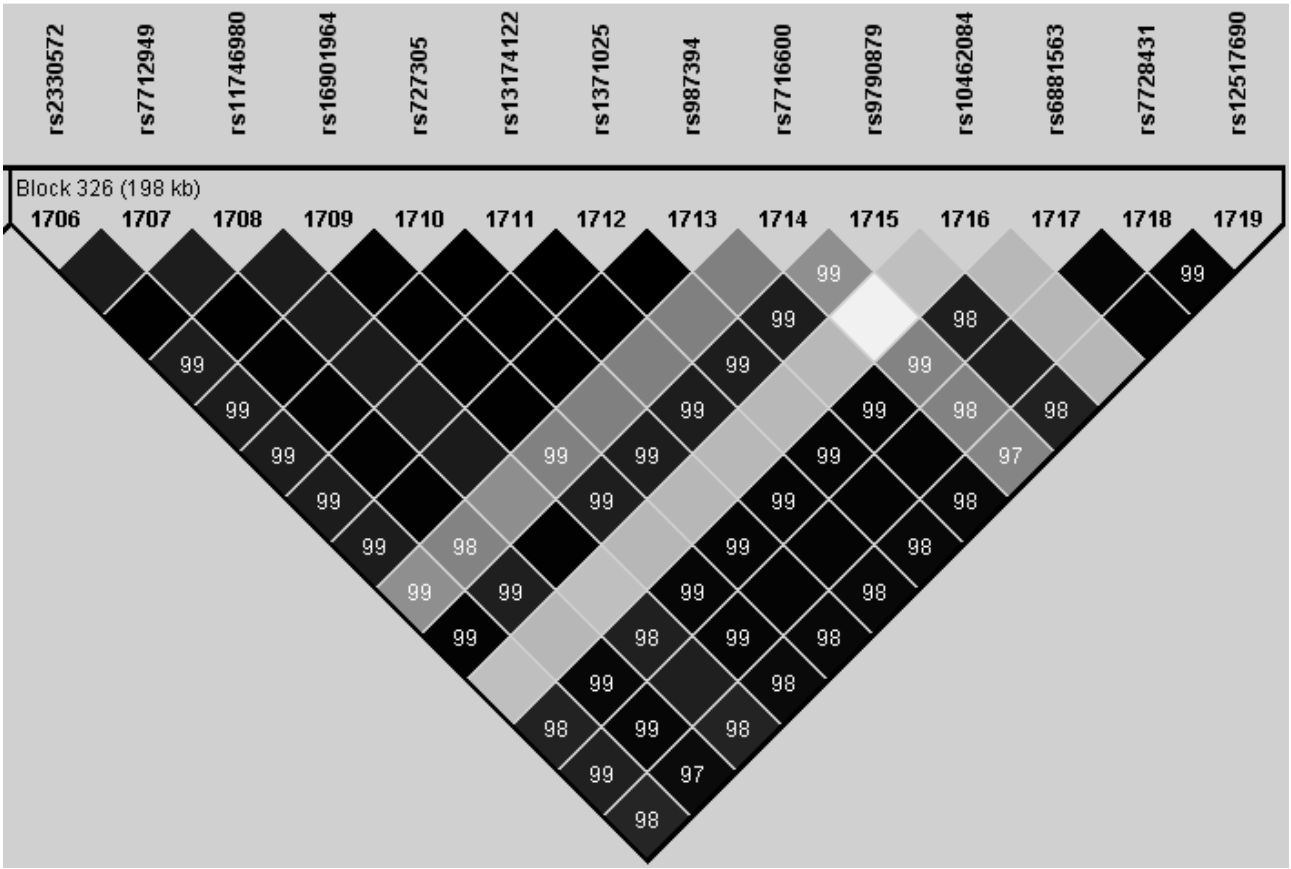
Watson, J. D. & Crick, F. H. C. (1953), "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid", *Nature* (Vol. 171, no. 4356), 737–738, Macmillan Journals Limited, Iso-Britannia.

Liitteet

Liite 1. χ^2-testin 15 merkittävintä SNP-markkeria.

Kromosomi	SNP	lokus	A1-emäs	A2-emäs	χ^2-testisuure	p-arvo
17	rs8066558	27296992	A	C	166,6	4,17E-38
10	rs7905923	109663117	G	A	22,9	1,70E-06
1	rs12123980	8594533	A	C	22,8	1,80E-06
19	rs4807425	3180224	A	G	22,64	1,96E-06
17	rs319749	28596020	A	G	20,68	5,44E-06
10	rs4255475	48144906	A	C	20,4	6,27E-06
10	rs7915527	109721574	A	G	20,04	7,58E-06
18	rs965174	54995424	A	G	19,04	1,28E-05
9	rs7046415	80670516	G	A	17,91	2,31E-05
21	rs2834950	35803265	A	G	17,35	3,11E-05
4	rs11937144	5927924	A	G	17,26	3,26E-05
21	rs2834650	35131970	A	G	17,2	3,37E-05
11	rs541207	63881718	C	A	16,98	3,78E-05
9	rs12336488	9218107	G	A	16,64	4,53E-05
2	rs10193128	233695966	A	G	16,57	4,68E-05

Liite 2. Kromosomin 5 haplotyyppiblokki, jossa on 14 SNP-markkeria.



Liite 3. Kromosomi 11 haplotyyppiblokkien p-arvokuvaaja.